Probabilistic Optimal Estimation and Filtering under Uncertainty

FABRIZIO DABBENE, MARIO SZNAIER, AND ROBERTO TEMPO*

Abstract

The classical approach to system identification is based on statistical assumptions about the measurement error, and provides estimates that have stochastic nature. Worst-case identification, on the other hand, only assumes the knowledge of deterministic error bounds, and provides guaranteed estimates, thus being in principle better suited for its use in control design. However, a main limitation of such deterministic bounds lies on the fact that they often turn out to be overly conservative, thus leading to estimates of limited use.

In this paper, we propose a rapprochement between these two paradigms, stochastic and worst-case, and propose a novel probabilistic framework for system identification that combines elements from information-based complexity with recent developments in the theory of randomized algorithms. The main idea in this line of research is to "discard" sets of measure at most ϵ, where ϵ is a probabilistic accuracy, from the set of deterministic estimates. Therefore, we are decreasing the so-called worst-case radius of information at the expense of a given probabilistic "risk."

In this setting, we compute a trade-off curve, called *violation function*, which shows how the radius of information decreases as a function of the accuracy. To this end, we construct randomized and deterministic algorithms which provide approximations of this function. The obtained results are based upon specific properties regarding the intersection of convex sets.

Keywords: System identification and filtering, optimal algorithms, randomized algorithms, uncertain systems

* Corresponding author

F. Dabbene and R. Tempo are with the CNR-IEIIT Institute, Politecnico di Torino, Italy (e-mail: fabrizio.dabbene@polito.it, roberto.tempo@polito.it).

M. Sznaier is with the Northeastern University, Boston, MA 02115, USA (e-mail: msznaier@ece.neu.edu).

Preliminarily versions of these results have been submitted for conference publication, see [16], [17].

I. Introduction and Preliminaries

Information-based complexity (IBC) is a general theory developed within computer science for studying the complexity of problems approximately solved due to the presence of partial and/or contaminated information; see [52] and [53] and references therein. Typical applications of IBC include distributed computations, clock synchronization, computer vision, numerical integration, solution of nonlinear equations and large scale linear systems, as well as system identification and filtering. Depending on the specific application, different settings are studied: worst-case, average-case and probabilistic. In each setting, the objective is threefold: to derive optimal algorithms, to compute the associated approximation errors and to establish their computational complexity.

In the last decades, several authors focused their attention on the so-called set-membership identification which aims at the computation of hard bounds on the estimation errors, see for instance [36], and [26] for pointers to more recent developments. Set-membership identification may be embedded within the general framework of worst-case IBC, so that various systems and control problems, such as time-series analysis, filtering and $\mathcal{H}_\infty$ identification can be addressed [35], [49], [25], [3], [41]. In this setting, the noise is a deterministic variable bounded within a set of given radius. The objective is to derive optimal algorithms which minimize (with respect to the noise) the maximal distance between the true-but-unknown system parameters and their estimates. The drawback of the hard bound approach is that the estimation errors may be too large in many instances and therefore of limited use, in particular when the ultimate objective is to use system identification in the context of closed-loop control.

The worst-case setting provides a counterpart to the mainstream stochastic approach to system identification, see [33] and the special issues [34], [45], which has been very successful also in many applications, such as e.g. process control, systems biology and optimization. This approach assumes that the available observations are contaminated by *random* noise, and has the goal to derive soft bounds on the estimation errors. In this case, optimality is guaranteed in a probabilistic sense and the resulting algorithms often enjoy convergence properties only asymptotically.

The worst-case setting is based on the "concern" that the noise may be very malicious. The computed bounds are certainly more pessimistic than the stochastic ones, but the idea is to guard against the worst-case scenario, even though it is unlikely to occur. These observations lead us to discuss the *rapprochement* viewpoint, see [38], [23], [40], [12], [21], which has the following starting point: the measurement noise is confined within a given set (and therefore it falls under the framework of the worst-case setting), but it is also a random variable with given probability distribution (so that statistical information is used). A simple example is uniformly distributed noise with a supporting set which is that adopted by the worst-case methods. We recall that the rapprochement approach has

been extensively studied in the context of control design in the presence of uncertainty, see [50], [10], [9], and [7]. This research provides a rigorous methodology for deriving controllers guaranteeing the desired performance specifications with high level of probability.

The focal point of this paper is to address the rapprochement between soft and hard bounds in a rigorous fashion, with the goal to derive useful computational tools for system identification and filtering. The specific problem formulation we consider is the probabilistic setting of information-based complexity and the objective is to compute (by means of randomized and deterministic algorithms) the so-called *probabilistic radius of information*, with emphasis on uniformly distributed noise. We remark that, contrary to the statistical setting which mainly concentrates on asymptotic results, the probabilistic radius introduced in this paper provides a quantification of the estimation error which is based on a finite number of observations. In this sense, this approach has close relations with the works based on statistical learning theory proposed in [29], [56], [54], and with the approach in [14], [15], where noise-free non-asymptotic confidence sets for the estimates are derived. Furthermore, the paper is also related to the work [51], where a probabilistic density function over the consistency set is considered.

We now provide a preview of the structure and main results of the paper. Section III presents an introduction to information-based complexity. To summarize, we are interested in computing an optimal approximation of $\mathcal{S}x \in Z$ where $\mathcal{S}$ is a given linear mapping $\mathcal{S} : X \to Z \subseteq \mathbb{R}^s$, and $x \in X \subseteq \mathbb{R}^n$; x and $\mathcal{S}$ are called, respectively, problem element and solution operator (that is, $\mathcal{S}x$ represents a linear combination of the unknown parameters of the system to be identified). The element x is not exactly known but only approximate information $y = \mathcal{I}x + q$ is available, where $\mathcal{I}$ is called information operator, which is linear, and the noise q is confined within a norm-bounded set $\mathcal{Q} \subset \mathbb{R}^m$. An approximation to $\mathcal{S}x$ is provided by an algorithm (or estimator) $\mathcal{A}$, generally nonlinear, acting on the information y. Optimal algorithms minimize the maximal distance between the true-but-unknown solution $\mathcal{S}x$ and the estimated solution $\mathcal{A}(y)$ for the worst-case noise $q \in \mathcal{Q}$. The error of an optimal algorithm is called the worst-case radius of information. This section also contains the formal definition of the consistency set $\mathcal{I}^{-1}(y)$ which plays a major role in this paper. Roughly speaking, this is the set of all parameters x which are compatible with the given data y, the model $y = \mathcal{I}x + q$ and the noise $q \in \mathcal{Q}$. Section III includes two examples demonstrating how system parameter identification, prediction and filtering may be formulated in the general IBC framework.

Section IV introduces the probabilistic setting. In this context, the idea is to "discard" sets of (probabilistic) measure at most ϵ from the consistency set, having the objective of decreasing the worst-case radius, thus obtaining a new error which represents the probabilistic radius of information. In other words, the worst-case radius is decreased, possibly significantly, at the expense of a probabilistic risk ϵ. To discard "bad sets" of small measure is exactly the main feature of the probabilistic approach

to analysis and design of uncertain systems proposed in [50]. This approach may be very useful, for example, for system identification in the presence of outliers [2], where "bad measurements" may be discarded. In this section, by means of a chance-constrained approach [37], we also show that the probabilistic radius is related to the minimization of the so-called *optimal violation function* $v_{\mathrm{o}}(r)$. Roughly speaking, this function describes how the risk ϵ decreases as a function of the radius r. However, the computation of $v_o(r)$ is not an easy task and requires the results proved in Section V and the algorithms presented in Section VI. Section IV ends showing a tutorial example regarding estimation of the parameters of a second order model corrupted by additive noise. The example in continued in other sections of the paper for illustrative purposes.

Section V deals with uniformly distributed noise and contains the main technical results of the paper. Theorem 1 shows that the induced measure (through the inverse of the operator information $\mathcal{I}$) over the consistency set $\mathcal{I}^{-1}(y)$ is uniform and the induced measure of the set $\mathcal{S}\mathcal{I}^{-1}(y)$ (which is the transformation of the consistency set $\mathcal{I}^{-1}(y)$ through the solution operator $\mathcal{S}$) is log-concave. Theorem 2 proves crucial properties, from the computational point of view, of the optimal violation function $v_o(r)$. In particular, this result shows that $v_{\mathrm{o}}(r)$ is non-increasing, and for fixed $r > 0$, it can be obtained as the maximization of a specially constructed unimodal function. Hence, it may be easily computed by means of various optimization techniques which are discussed in the next section.

In Section VI we introduce specific algorithms for computing the optimal violation function. First, we observe that the exact computation of $v_{\mathrm{o}}(r)$ requires the evaluation of the volume of polytopes. Since this problem is NP-Hard [30], we propose to use suitable probabilistic and deterministic relaxations. More precisely, first we present a randomized algorithm based upon the classical Markov Chain Monte Carlo method [48], [47], which has been studied in the context of randomized stochastic approximation methods [32]; see also [50] and [9] for further details about randomized algorithms. Secondly, we present a deterministic relaxation of $v_{\mathrm{o}}(r)$ which is based upon the solution of a semi-definite program (SDP). The performance of both algorithms is shown using the example given in Section V.

Section VII discusses normally distributed noise, and presents some connections with classical stochastic estimation. In particular, it is shown that the least-squares algorithm is optimal also in the probabilistic setting discussed in this paper. For this case, we state a bound (which is essentially tight for small-variance noise) on the probabilistic radius of information, which is given in [52] in terms of the so-called *average radius of information*, and it depends on ϵ, on the noise covariance, and on the information and solution operators.

II. NOTATION

In this section, we provide the notation used in this paper. We write $\|\cdot\|$, $\|\cdot\|_2$ and $\|\cdot\|_\infty$ to denote the ℓ_p, ℓ_2 and ℓ_∞ norms, respectively. The ℓ_p norm-ball of center ξ_c and radius r is defined as

$$\mathcal{B}(\xi_c, r) \doteq \{\xi \,|\, \|\xi - \xi_c\| \leq r\},$$

and we write $\mathcal{B}(r) \doteq \mathcal{B}(0, r)$. We denote by $\mathcal{B}_2(\xi_c, r)$ and by $\mathcal{B}_\infty(\xi_c, r)$ the ℓ_2 and ℓ_∞ norm-balls, respectively. We use the notation $x \sim p_A$ to indicate that the random vector x has probability density function (pdf) $p_A(x)$ with support set A. The *uniform* density $\mathcal{U}_A$ over the set $A \subset \mathbb{R}^n$ is defined as

$$\mathcal{U}_A(x) \doteq \begin{cases} 1/\mathrm{vol}\,[A] & \text{if } x \in A; \\ 0 & \text{otherwise} \end{cases}$$

where $\mathrm{vol}\,[A]$ represents the Lebesgue measure (volume) of the set A, see [22] for details regarding volumes of sets. The uniform density $\mathcal{U}_A$ generates a uniform measure $\mu_{\mathcal{U}(A)}$ such that, for any measurable set B, $\mu_{\mathcal{U}(A)}(B) = \mathrm{vol}\,[B \cap A]\,/\mathrm{vol}\,[A]$. The multivariate *normal density* $\mathcal{N}_{\bar{x},\Sigma}$ with mean value $\bar{x} \in \mathbb{R}^n$, symmetric positive definite covariance matrix $\Sigma \succ 0$ is defined as

$$\mathcal{N}_{\bar{x},\Sigma}(x) \doteq (2\pi)^{-n/2}|\Sigma|^{-1/2}\,\mathrm{e}^{-\frac{1}{2}(x-\bar{x})^\top\Sigma^{-1}(x-\bar{x})}.$$

The normal probability density function generates the so-called Gaussian measure $\mu_{\mathcal{N}}$. The (univariate) *unilateral Gamma density* with parameters $a, b \in \mathbb{R}$ is defined

$$G_{a,b}(x) = \frac{1}{\Gamma(a)b^a}x^{a-1}\mathrm{e}^{-x/b}, \quad x \geq 0, \tag{1}$$

where $\Gamma(\cdot)$ is the Gamma function. We denote by $\mathbb{I}\,(\cdot)$ the indicator function, which is equal to one if the clause is true, and it is zero otherwise. The $n \times n$ identity and the $n \times m$ zero matrices are indicated by I_n and $0_{n,m}$, respectively. A set H is said to be *centrally symmetric* with center $\bar{x}$ if $x \in H$ implies that its reflection with respect to $\bar{x}$ also belongs to H, i.e. $(2\bar{x} - x) \in H$.

III. INFORMATION-BASED COMPLEXITY

This section provides a formal overview of the information-based complexity definitions used in the paper and two introductory examples illustrating the main concepts in a system identification and filtering contexts. An introduction to the IBC framework is given in [53] and in [41]; see the monograph [52] for an advanced treatment of the topic. The relevant spaces, operators and sets discussed next are shown in Figure 1.

Let X be a linear normed n-dimensional space over the real field, which represents the set of (unknown) problem elements $x \in X$. Define a linear operator $\mathcal{I}$, called *information operator*, which maps X into a linear normed m-dimensional space Y

$$\mathcal{I} : X \to Y.$$

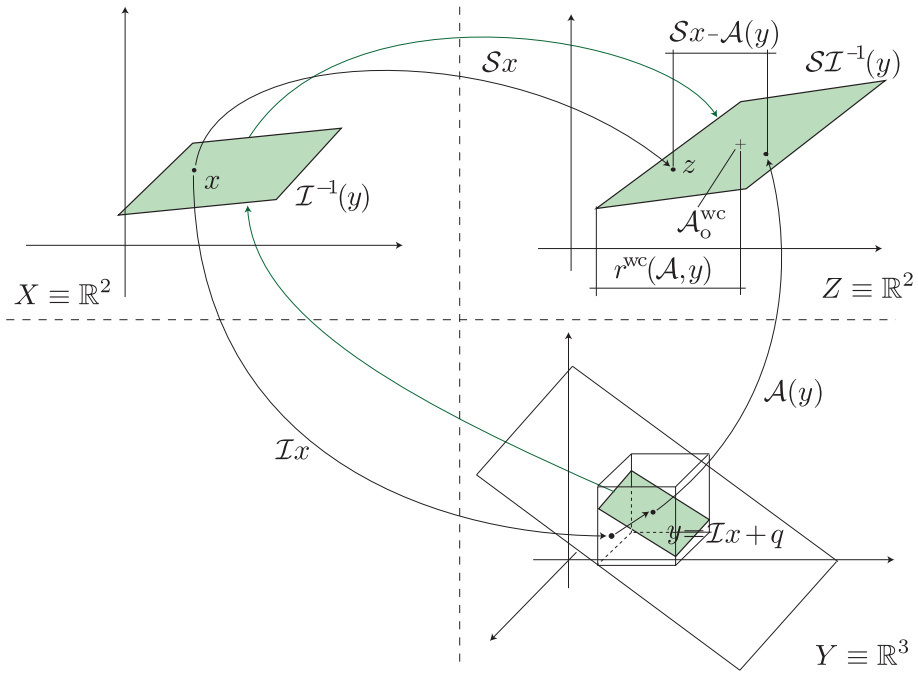

Fig. 1. Illustration of the information-based complexity framework.

In general, exact information about the problem element $x \in X$ is not available and only perturbed information, or *data*, $y \in Y$ is given. That is, we have

$$y = \mathcal{I}x + q \tag{2}$$

where q represents *uncertainty* which may be deterministic or random. In either case, we assume that $q \in \mathcal{Q}$, where $\mathcal{Q} \subseteq \mathbb{R}^m$ is a bounding set. Due to the presence of uncertainty q, the problem element $x \in X$ may not be easily recovered knowing data $y \in Y$. Then, we introduce a linear operator $\mathcal{S}$, called a *solution operator*, which maps X into Z

$$\mathcal{S} : X \to Z$$

where Z is a linear normed s-dimensional space over the real field, where $s \leq n$. Given $\mathcal{S}$, our aim is to estimate an element $\mathcal{S}x \in Z$ knowing the corrupted information $y \in Y$ about the problem element $x \in X$.

An *algorithm* $\mathcal{A}$ is a mapping (in general nonlinear) from Y into Z, i.e.

$$\mathcal{A} : Y \to Z.$$

An algorithm provides an approximation $\mathcal{A}(y)$ of $\mathcal{S}x$ using the available information $y \in Y$ of $x \in X$. The outcome of such an algorithm is called an *estimator* $z = \mathcal{A}(y)$.

We now introduce a set which plays a key role in the subsequent definitions of radius of information and optimal algorithm. Given data $y \in Y$, we define the *consistency set* as follows

$$\mathcal{I}^{-1}(y) \doteq \{x \in X \mid \text{ there exists } q \in \mathcal{Q} : y = \mathcal{I}x + q\} \tag{3}$$

which represents the set of all problem elements $x \in X$ compatible with (i.e. not invalidated by) $\mathcal{I}x$, uncertainty q and bounding set $\mathcal{Q}$. Note that, under the sufficient information assumption stated next, the set $\mathcal{I}^{-1}(y)$ is guaranteed to be bounded. For the sake of simplicity, we assume that the three sets X, Y, Z are equipped by the same ℓ_p norm, and that the set $\mathcal{Q}$ is an ℓ_p norm-ball of radius ρ, that is $\mathcal{Q} \equiv \mathcal{B}(\rho)$. Note that in this case the set $\mathcal{I}^{-1}(y)$ can be written as

$$\mathcal{I}^{-1}(y) = \{x \in X \mid \|\mathcal{I}x - y\| \leq \rho\}. \tag{4}$$

The following assumption regarding the operators $\mathcal{I}$ and $\mathcal{S}$ is now introduced.

Assumption 1 (Sufficient information and feasibility): We assume that the information operator $\mathcal{I}$ is a one-to-one mapping, i.e. $m \geq n$ and $\operatorname{rank} \mathcal{I} = n$. Similarly, $n \geq s$ and $\mathcal{S}$ is full row rank. Moreover, we assume that the set $\mathcal{I}^{-1}(y)$ has non-empty interior.

Note that, in a system identification context, the assumption on $\mathcal{I}$ and on the consistency set are necessary conditions for identifiability of the problem element $x \in X$. The assumption of full-rank $\mathcal{S}$ is equivalent to assuming that the elements of the vector $z = \mathcal{S}x$ are linearly independent (otherwise, one could always estimate a linearly independent set and use it to reconstruct the rest of the vector z). We now provide an illustrative example showing the role of these operators and spaces in the context of system identification.

Example 1 (System parameter identification and prediction): Consider a parameter identification problem which has the objective to identify a linear system from noisy measurements. In this case, the problem elements are represented by the trajectory $\xi = \xi(t, x)$ of a dynamic system, parametrized by some unknown parameter vector $x \in X$. This may for instance be represented as follows

$$\xi(t, x) = \sum_{i=1}^{n} x_i \psi_i(t) = \Psi^\top(t) x,$$

with given basis functions $\psi_i(t)$, and $\Psi^\top(t) \doteq [\psi_1(t) \quad \cdots \quad \psi_n(t)]$. We suppose that m noisy measurements of $\xi(t, x)$ are available for $t_1 < t_2 < \cdots < t_m$, that is

$$y = \mathcal{I}x + q = [\Psi(t_1) \quad \cdots \quad \Psi(t_m)]^\top x + q. \tag{5}$$

In this context, one usually assumes unknown but bounded errors, such that $|q_i| \leq \rho$, $i = 1, \ldots, m$, that is $\mathcal{Q} = \mathcal{B}_\infty(\rho)$. Then, the aim is to obtain a parameter estimate using the data y. Hence, the solution operator is given by the identity,

$$\mathcal{S}x = x$$

and $Z \equiv X$. The consistency set is sometimes referred to as feasible parameters set, and is given as follows

$$\mathcal{I}^{-1}(y) = \left\{ x \in X : \|y - [\Psi(t_1) \quad \cdots \quad \Psi(t_m)]^\top x\|_\infty \leq \rho \right\}. \tag{6}$$

In the case of time series prediction, we are interested on predicting s future values of the function $\xi(t,x)$ based on m past measurements, and the solution operator takes the form

$$z = \mathcal{S}x = \xi(t_{m+1}, x) = \Psi^\top(t_{m+1})x$$

for a one-step prediction, and by

$$z = \mathcal{S}x = \{\xi(t_{m+1}, x), \quad \cdots \quad , \xi(t_{m+s}, x)\} = [\Psi(t_{m+1}) \quad \cdots \quad \Psi(t_{m+s})]^\top x,$$

for a s steps prediction.

Example 2 (Filtering of dynamic systems): Note that the IBC theoretical setting outlined so-far also applies to the problem of filtering. Consider an LTI system described by the following state-space model:

$$\begin{aligned} \xi_{k+1} &= A\xi_k + B_k w_k, \ w_k \in \mathcal{W}; \\ y_k &= C_k \xi_k + q_k, \ q_k \in \mathcal{Q} \\ z_k &= H_k \xi_k \end{aligned} \tag{7}$$

where w_k and q_k denote process and measurement noise characterized by a set-membership description given by the compact, convex sets $\mathcal{W}$ and $\mathcal{Q}$. The goal is to estimate z_n, the value of the variable z at a given time instant n from the noisy measurements $y_k, \ k = 0, \ldots, n$. We will further assume that the pairs (A, B) and (A, C) are controllable and observable, respectively, and that rank$(H) = s$, that is, the elements of z are linearly independent This problem fits the formalism described above by defining

$$x \doteq \begin{bmatrix} \xi_o \\ w_o \\ \vdots \\ w_{n-1} \end{bmatrix}, \ y \doteq \begin{bmatrix} y_o \\ y_1 \\ \vdots \\ y_n \end{bmatrix}$$

$$\mathcal{I} = \begin{bmatrix} C & 0 & 0 & 0 & \ldots & 0 \\ CA & CB & 0 & 0 & \ldots & 0 \\ CA^2 & CAB & CB & 0 & \ldots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ CA^n & CA^{n-1}B & CA^{n-2}B & \ldots & CAB & CB \end{bmatrix} \tag{8}$$

$$\mathcal{S} \doteq \begin{bmatrix} HA^n & HA^{n-1}B & HA^{n-2}B & \ldots & HAB & HB \end{bmatrix}$$

Note that in the formulation above, observability of the pair (A, C) guarantees that $\mathcal{I}$ has rank n, while the controllability of (A, B), coupled with rank$(H) = s$ guarantees that $\mathcal{S}$ has full row rank.

The case where $\mathcal{W}$ and $\mathcal{Q}$ are ℓ_∞ balls in the respective spaces was addressed in [35] and [49]. The latter showed that a worst-case globally optimal estimator is linear in the measurements and can be found based on solving a single LP problem. However, this estimator is non-recursive and its complexity cannot be bounded a priori. Related results were also discussed in [55], [43].

Next, we define approximation errors and optimal algorithms when q is deterministic or random. First, we briefly summarize the deterministic case which has been deeply analyzed in the literature, see e.g. [35]. The definitions concerning the probabilistic case are new in this context, and are introduced in Section IV.

A. Worst-Case Setting

Given data $y \in Y$, we define the worst-case error $r^{\mathrm{wc}}(\mathcal{A}, y)$ of the algorithm $\mathcal{A}$ as

$$r^{\mathrm{wc}}(\mathcal{A}, y) \doteq \max_{x \in \mathcal{I}^{-1}(y)} \|\mathcal{S}x - \mathcal{A}(y)\|. \tag{9}$$

This error is based on the available information $y \in Y$ about the problem element $x \in X$ and it measures the approximation error between Sx and $A(y)$. An algorithm $\mathcal{A}_{\mathrm{o}}^{\mathrm{wc}}$ is called *worst-case optimal* if it minimizes $r^{\mathrm{wc}}(\mathcal{A}, y)$ for any $y \in Y$. That is, given data $y \in Y$, we have

$$r_{\mathrm{o}}^{\mathrm{wc}}(y) \doteq r^{\mathrm{wc}}(\mathcal{A}_{\mathrm{o}}^{\mathrm{wc}}, y) \doteq \inf_{\mathcal{A}} r^{\mathrm{wc}}(\mathcal{A}, y). \tag{10}$$

The minimal error $r_{\mathrm{o}}^{\mathrm{wc}}(y)$ is called the *worst-case radius of information*[1].

This optimality criterion is meaningful in estimation problems as it ensures the smallest approximation error between the actual (unknown) solution $\mathcal{S}x$ and its estimate $\mathcal{A}(y)$ for the worst element $x \in \mathcal{I}^{-1}(y)$ for any given data $y \in Y$. Obviously, a worst-case optimal estimator is given by $z_{\mathrm{o}}^{\mathrm{wc}} = \mathcal{A}_{\mathrm{o}}^{\mathrm{wc}}(y)$, see Figure 1.

We notice that optimal algorithms map data y into the *ℓ_p–Chebychev center* of the set $\mathcal{S}\mathcal{I}^{-1}(y)$, where the Chebychev center $z_{\mathrm{c}}(H)$ of a set $H \subseteq Z$ is defined as

$$\max_{h \in H} \|h - z_{\mathrm{c}}(H)\| \doteq \inf_{z \in Z} \max_{h \in H} \|h - z\| \doteq r_{\mathrm{c}}(H).$$

Optimal algorithms are often called *central algorithms* and $z_{\mathrm{c}}(\mathcal{S}\mathcal{I}^{-1}(y)) = z_{\mathrm{o}}^{\mathrm{wc}}$. We remark that, in general, the Chebychev center of a set $H \subset Z$ may not be unique and not necessarily belongs to H, for example, when H is not convex or it is a discrete set.

[1]In the IBC context, this error is usually referred to as "local" radius of information, to distinguish it from the so-called "global" radius, see [52] for further details.

Remark 1 (Geometric interpretation of the Chebychev center): Note that, by construction, for any given set H (not necessarily convex, nor connected), the ℓ_p–Chebychev center $z_{\rm c}(H)$ of H and its radius $r_{\rm c}(H)$ are given by the center and radius of the smallest ℓ_p norm-ball enclosing the set H, see Figure 2.

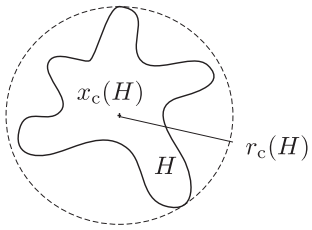

Fig. 2. Chebychev center and radius of a generic set H.

That is, we can compute $z_{\rm c}(H)$ and $r_{\rm c}(H)$ solving the optimization problem

$$\min_{z,r} r \quad \text{subject to } \mathcal{B}(z,r) \supseteq H. \tag{11}$$

Note that, as remarked above, the optimal ball $\mathcal{B}(z,r)$ need not be unique. It follows immediately that if the set H is centrally symmetric with center $\bar{z}$, then $\bar{z}$ is a Chebychev center of H. $\diamond$

The computation of the worst-case radius of information $r_{\rm o}^{\rm wc}(y)$ and of the derivation of optimal algorithms $\mathcal{A}_{\rm o}^{\rm wc}$ has been the focal point of several papers in a system identification setting. For example, for ℓ_∞ norms in equation (9), it has been shown in [35] that an optimal algorithm and the radius of information can be computed solving $2n$ linear programs. In particular, an optimal algorithm is easily obtained computing the center of the tightest hyperrectangle which contains the polytope $\mathcal{S}\mathcal{I}^{-1}(y)$. If the norm is ℓ_2, then $\mathcal{S}\mathcal{I}^{-1}(y)$ is an ellipsoid, and the linear optimal estimator is the least squares algorithm

$$\mathcal{A}_{\rm ls}(y) = \mathcal{S}x_{\rm ls} \tag{12}$$

where $x_{\rm ls}$ is defined as

$$\|\mathcal{I}x_{\rm ls} - y\|_2 \doteq \min_{x\in X} \|\mathcal{I}x - y\|_2.$$

In this case, see e.g. [27], $\mathcal{A}_{\rm ls}(y)$ is the Chebychev center of the ellipsoid $\mathcal{S}\mathcal{I}^{-1}(y)$,

$$z_{\rm o}^{\rm wc} = \mathcal{A}_{\rm ls}(y) = \mathcal{S}(\mathcal{I}^\top\mathcal{I})^{-1}\mathcal{I}^\top y$$

and its worst-case error is given by

$$r_{\rm o}^{\rm wc}(y) = r^{\rm wc}(\mathcal{A}_{\rm ls}, y) = \rho\sqrt{\lambda_{\max}\left(\mathcal{S}^\top(\mathcal{I}^\top\mathcal{I})^{-1}\mathcal{S}\right)}$$

where ρ is defined in (4). In general, $\mathcal{A}_{\rm ls}$ is not optimal for other bounding sets such as ℓ_∞ balls.

IV. Probabilistic Setting with Random Uncertainty

In this section, we introduce a probabilistic counterpart of the worst-case setting previously defined, that is we define optimal algorithms $\mathcal{A}_{\mathrm{o}}^{\mathrm{pr}}$ and the probabilistic radius $r^{\mathrm{pr}}(\mathcal{A}, y, \epsilon)$ for the so-called probabilistic setting when the uncertainty q is random. and $\epsilon \in (0,\, 1)$ is a given parameter called *accuracy*. Roughly speaking, in this setting the error of an algorithm is measured in a worst-case sense, but we "discard" a set of measure at most ϵ from the consistency set $\mathcal{SI}^{-1}(y)$. Hence, the probabilistic radius of information may be interpreted as the smallest radius of a ball discarding a set whose measure is at most ϵ, see Figure 3. Therefore, we are decreasing the worst-case radius of information at the expense of a probabilistic "risk" ϵ. In a system identification context, reducing the radius of information is clearly a highly desirable property. Using this probabilistic notion, we can compute a trade-off function which shows how the radius of information decreases as a function of the parameter ϵ, as described in the tutorial Example 3 and in the numerical example presented in Section VIII.

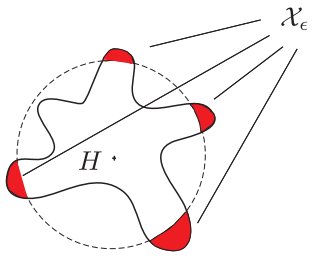

Fig. 3. Illustration of the probabilistic framework for a generic set H. The measure of the dark (red) set $\mathcal{X}_\epsilon$ – discarded from the consistency set – is at most ϵ, and the probabilistic radius is smaller than the worst-case radius.

We now state a formal assumption regarding the random uncertainty q.

Assumption 2 (Random measurement uncertainty): We assume that the uncertainty q is a real random[2] vector with given probability density $p_{\mathcal{Q}}(q)$ and support set $\mathcal{Q} = \mathcal{B}(\rho)$. We denote by $\mu_{\mathcal{Q}}$ the probability measure generated by $p_{\mathcal{Q}}(q)$ over the set $\mathcal{Q}$.

Remark 2 (Induced measure over $\mathcal{I}^{-1}(y)$): We note that the probability measure over the set $\mathcal{Q}$ *induces*, by means of equation (2), a probability measure $\tilde{\mu}_{\mathcal{I}^{-1}}$ over the set $\mathcal{I}^{-1}(y)$. Indeed, for any measurable set $B \subseteq X$, we consider the probability measure $\mu_{\mathcal{Q}}$ as follows: $\tilde{\mu}_{\mathcal{I}^{-1}}(B) = \mu_Q(q \in \mathcal{Q} \,|\, \exists x \in B \cap \mathcal{I}^{-1}(y) \,|\, \mathcal{I}x + q = y)$. This probability measure is such that points outside the consistency set $\mathcal{I}^{-1}(y)$ have measure zero, and $\tilde{\mu}_{\mathcal{I}^{-1}}\big(\mathcal{I}^{-1}(y)\big) = 1$, that is this induced measure is concentrated over $\mathcal{I}^{-1}(y)$. This induced measure is formally defined in [52, Chapter 6], where it is shown that it

[2]For simplicity, in this assumption we consider the case when the density of q exists (that is the distribution is differentiable).

is indeed a *conditional measure*. Similarly, we denote by $\tilde{p}_{\mathcal{I}^{-1}}$ the induced probability density, having support over $\mathcal{I}^{-1}(y)$. We remark that Theorem 1 in Section V studies the induced measure $\tilde{\mu}_{\mathcal{I}^{-1}}(\cdot)$ over the set $\mathcal{I}^{-1}(y)$ when q is uniformly distributed within $\mathcal{Q}$, showing that this measure is still uniform. In turn, the induced measure $\tilde{\mu}_{\mathcal{I}^{-1}}$ is mapped through the linear operator $\mathcal{S}$ into a measure over $\mathcal{S}\mathcal{I}^{-1}(y)$, which we denote as $\tilde{\mu}_{\mathcal{S}\mathcal{I}^{-1}}$. Similarly, the induced density is denoted as $\tilde{p}_{\mathcal{S}\mathcal{I}^{-1}}$. In Theorem 1 in Section V we show that the induced measure $\tilde{\mu}_{\mathcal{S}\mathcal{I}^{-1}}$ is log-concave in the case of uniform density over $\mathcal{Q}$. $\diamond$

Given corrupted information $y \in Y$ and accuracy $\epsilon \in (0,1)$, we define the probabilistic error (to level ϵ) $r^{\mathrm{pr}}(\mathcal{A}, y, \epsilon)$ of the algorithm $\mathcal{A}$ as

$$r^{\mathrm{pr}}(\mathcal{A}, y, \epsilon) \doteq \inf_{\mathcal{X}_\epsilon \text{ such that } \tilde{\mu}_{\mathcal{I}^{-1}}(\mathcal{X}_\epsilon) \leq \epsilon} \quad \max_{x \in \mathcal{I}^{-1}(y) \setminus \mathcal{X}_\epsilon} \|\mathcal{S}x - \mathcal{A}(y)\| \tag{13}$$

where the notation $\mathcal{I}^{-1}(y) \setminus \mathcal{X}_\epsilon$ indicates the set-theoretic difference between $\mathcal{I}^{-1}(y)$ and $\mathcal{X}_\epsilon$,

$$\mathcal{I}^{-1}(y) \setminus \mathcal{X}_\epsilon \doteq \left\{ x \in \mathcal{I}^{-1}(y) \,|\, x \notin \mathcal{X}_\epsilon \right\}.$$

Clearly, $r^{\mathrm{pr}}(\mathcal{A}, y, \epsilon) \leq r^{\mathrm{wc}}(\mathcal{A}, y)$ for any algorithm $\mathcal{A}$, data $y \in Y$ and accuracy level $\epsilon \in (0,1)$, which implies a reduction of the approximation error in a probabilistic setting.

An algorithm $\mathcal{A}_{\mathrm{o}}^{\mathrm{pr}}$ is called *probabilistic optimal* (to level ϵ) if it minimizes the error $r^{\mathrm{pr}}(\mathcal{A}, y, \epsilon)$ for any $y \in Y$ and $\epsilon \in (0,1)$. That is, given data $y \in Y$ and accuracy level $\epsilon \in (0,1)$, we have

$$r_{\mathrm{o}}^{\mathrm{pr}}(y, \epsilon) \doteq r^{\mathrm{pr}}(\mathcal{A}_{\mathrm{o}}^{\mathrm{pr}}, y, \epsilon) = \inf_{\mathcal{A}} r^{\mathrm{pr}}(\mathcal{A}, y, \epsilon). \tag{14}$$

The minimal error $r_{\mathrm{o}}^{\mathrm{pr}}(y, \epsilon)$ is called the *probabilistic radius of information* (to level ϵ) and the corresponding optimal estimator is given by

$$z_{\mathrm{o}}^{\mathrm{pr}}(\epsilon) \doteq \mathcal{A}_{\mathrm{o}}^{\mathrm{pr}}(y, \epsilon). \tag{15}$$

The problem we study in the next section is the computation of $r_{\mathrm{o}}^{\mathrm{pr}}(y, \epsilon)$ and the derivation of probabilistic optimal algorithms $\mathcal{A}_{\mathrm{o}}^{\mathrm{pr}}$. To this end, as in [52], we reformulate equation (13) in terms of a chance-constrained optimization problem [37]

$$r^{\mathrm{pr}}(\mathcal{A}, y, \epsilon) = \min \left\{ r \,|\, v(r, \mathcal{A}) \leq \epsilon \right\},$$

where the violation function for given algorithm $\mathcal{A}$ and radius r is defined as

$$v(r, \mathcal{A}) \doteq \tilde{\mu}_{\mathcal{I}^{-1}} \left\{ x \in \mathcal{I}^{-1}(y) \,|\, \|\mathcal{S}x - \mathcal{A}(y)\| > r \right\}.$$

Then, this formulation leads immediately to

$$r_{\mathrm{o}}^{\mathrm{pr}}(y, \epsilon) = \min \left\{ r \,|\, v_{\mathrm{o}}(r) \leq \epsilon \right\}, \tag{16}$$

where the *optimal violation function* for a given radius r is given by

$$v_{\mathrm{o}}(r) \doteq \inf_{\mathcal{A}} \tilde{\mu}_{\mathcal{I}^{-1}} \left\{ x \in \mathcal{I}^{-1}(y) : \|\mathcal{S}x - \mathcal{A}(y)\| > r \right\}. \tag{17}$$

To illustrate the notions introduced so far, we consider the following numerical example. The example is tutorial, and it is sufficiently simple so that all relevant sets are two dimensional and can be easily depicted.

Example 3 (Identification of a second order model): We consider a particular case of the system parameter identification problem introduced in Example 1. Our aim is to estimate the parameters of a second order model

$$y_k = x_1 u_k + x_2 u_{k-1} + q_k, \quad k = 1, \dots, m \tag{18}$$

where the input u_k is a known input sequence. The (unknown) nominal parameters were set to $[1.25 \quad 2.35]^\top$, and $m = 100$ measurements were collected generating the input sequence $\{u_k\}$ according to a Gaussian distribution with zero mean value and unit variance, and the measurement uncertainty q as a sequence of uniformly distributed noise with $|q_k| \leq 0.5$. Note that, in this simple case, the operator $\mathcal{S}$ is the identity, and thus $X \equiv Y$ and the sets $\mathcal{I}^{-1}(y)$ and $\mathcal{S}\mathcal{I}^{-1}(y)$ coincide. That is, the solution coincides with the parameters, and the goal is to estimate $z_i = x_i$, $i = 1, 2$.

First, the optimal worst-case radius defined in (10) and the corresponding optimal solution have been computed by solving four linear programs (corresponding to finding the tightest box containing the polytope $\mathcal{S}\mathcal{I}^{-1}(y)$). The computed worst-case optimal estimate is $z_{\rm o}^{\rm wc} = [1.2499 \quad 2.3551]^\top$ and the worst-case radius is $r_{\rm o}^{\rm wc}(y) = 0.0352$. For comparison, we also computed the classical least-squares estimate defined in (12) obtaining $\mathcal{A}_{\rm ls}(y) = [1.2873 \quad 2.3190]^\top$. Note that the least-square estimate is not worst-case optimal as expected, and it is not even *interpolatory*, since it is outside of the consistency set $\mathcal{S}\mathcal{I}^{-1}(y)$ (see [52] for a formal definition of interpolatory algorithm).

Subsequently, we fix the accuracy level $\epsilon = 0.1$, and aim at computing a probabilistic optimal radius and the corresponding optimal estimate according to definitions (14) and (15). By using the techniques discussed in Section V, we obtained $r_{\rm o}^{\rm pr}(y, 0.1) = 0.0284$ and $z_{\rm o}^{\rm pr}(0.1) = [1.2480 \quad 2.3540]^\top$, which represents a 25% improvement.

V. Random Uncertainty Uniformly Distributed

In this section, which contains the main technical results of the paper, we study the case when q is uniformly distributed over the ball $\mathcal{Q} \equiv \mathcal{B}(\rho)$, i.e. $q \sim \mathcal{U}_{\mathcal{Q}}$ and $\mu_{\mathcal{Q}} \equiv \mu_{\mathcal{U}(\mathcal{Q})}$. First, we address a preliminary technical question: *If $\mu_{\mathcal{Q}}$ is the uniform measure over $\mathcal{Q}$, what is the induced measure $\tilde{\mu}_{\mathcal{I}^{-1}}$ over the set $\mathcal{I}^{-1}(y)$ defined in equation (3)?* The next result shows that this distribution is indeed still uniform. Furthermore, we prove that the induced measure on $\mathcal{S}\mathcal{I}^{-1}(y)$ is log-concave.

Remark 3 (Log-concave measures and Brunn-Minkowski inequality): We recall that a measure $\mu(\cdot)$ is log-concave if, for any compact subsets A, B and $\lambda \in [0,\, 1]$, it holds

$$\mu(\lambda A + (1-\lambda)B) \geq \mu(A)^\lambda \mu(B)^{1-\lambda}$$

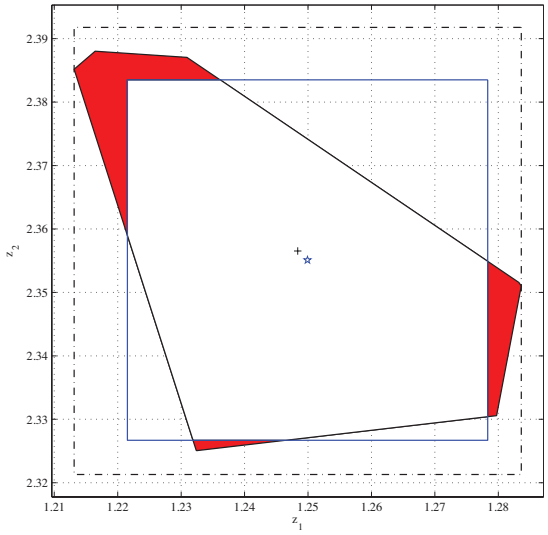

Fig. 4. Consistency set and relevant quantities for Example 3. The worst-case optimal radius $r_{\mathrm{o}}^{\mathrm{wc}}(y)$ corresponds to the radius of the dash-dotted box enclosing the polytope $\mathcal{S}\mathcal{I}^{-1}(y)$. Its center, denoted by a cross $+$, represents the optimal worst-case estimate $z_{\mathrm{o}}^{\mathrm{wc}}$. The probabilistic optimal (to level $\epsilon = 0.1$) radius $r_{\mathrm{o}}^{\mathrm{pr}}(y,\epsilon)$ corresponds to the radius of the solid-line box, which is the "optimal set" $\mathcal{X}_\epsilon$ that, according to definition (13), discards a set of measure ϵ from $\mathcal{I}^{-1}(y)$. The discarded set $\mathcal{I}^{-1}(y) \setminus \mathcal{X}_\epsilon$ is represented by the dark (red) area. The center of this box, denoted by a star $\star$, represents the optimal probabilistic estimate $z_{\mathrm{o}}^{\mathrm{pr}}(\epsilon)$.

where $\lambda A + (1-\lambda)B$ denotes the Minkowski sum[3] of the two sets λA and $(1-\lambda)B$. Note that the Brunn-Minkowski inequality [42] asserts that the uniform measure over convex sets is log-concave. Furthermore, any Gaussian measure is log-concave. $\diamond$

Theorem 1 (Measures over $\mathcal{I}^{-1}(y)$ and $\mathcal{S}\mathcal{I}^{-1}(y)$): Let $q \sim \mathcal{U}(\mathcal{Q})$ with $\mathcal{Q} \equiv \mathcal{B}(\rho)$, then, for any $y \in Y$ it holds:

(i) The induced measure $\tilde{\mu}_{\mathcal{I}^{-1}}$ is uniform over $\mathcal{I}^{-1}(y)$, that is $\tilde{\mu}_{\mathcal{I}^{-1}} \equiv \mu_{\mathcal{U}(\mathcal{I}^{-1}(y))}$;

(ii) The induced measure $\tilde{\mu}_{\mathcal{S}\mathcal{I}^{-1}}$ over $\mathcal{S}\mathcal{I}^{-1}(y)$ is log-concave. Moreover, if $\mathcal{S} \in \mathbb{R}^{n,n}$, then this measure is uniform, that is $\tilde{\mu}_{\mathcal{S}\mathcal{I}^{-1}} \equiv \mu_{\mathcal{U}(\mathcal{S}\mathcal{I}^{-1}(y))}$.

Proof: Consider the transformation matrix $T = [T_1\ T_2]$, where T_1 is an orthonormal basis of the column space of $\mathcal{I}$ and T_2 is an orthonormal basis of the null space of $\mathcal{I}^\top$. Furthermore, define the linear transformation $\bar{q} \doteq T^\top q$. Then, if we define the set $\bar{\mathcal{Q}} \doteq \left\{\bar{q} \in \mathbb{R}^m \,|\, T^{-\top}\bar{q} \in \mathcal{Q}\right\}$ we have that, if the random variable q is uniform on $\mathcal{Q}$, then the linearly transformed random variable $\bar{q}$ is uniform on $\bar{\mathcal{Q}}$ (see e.g. [22]). Then, by multiplying equation (2) from the left by $T^\top$, and defining $\bar{\mathcal{I}}_1 \doteq T_1^\top \mathcal{I}$

[3] The Minkowski sum of two sets A and B is obtained adding every element of A to every element of B, i.e. $A + B = \{a + b \,|\, a \in A, b \in B\}$.

and $\bar{y}_i \doteq T_i^\top y$, $\bar{q}_i \doteq T_i^\top q$, $i = 1, 2$, we get

$$\begin{cases} \bar{\mathcal{I}}_1 x & + & \bar{q}_1 & = \bar{y}_1 \\ & & \bar{q}_2 & = \bar{y}_2, \end{cases} \tag{19}$$

since, by construction, $T_2^\top \mathcal{I} = 0$. It follows from definition (3) that a point $x \in X$ belongs to $\mathcal{I}^{-1}(y)$ if and only if there exists $\bar{q} = \left[\begin{smallmatrix} \bar{q}_1 \\ \bar{q}_2 \end{smallmatrix}\right] \in \bar{\mathcal{Q}}$ such that (19) holds, i.e. if there exist $\bar{q}_1$ in the set $\bar{Q}_1 \doteq \left\{\bar{q}_1 \in \mathbb{R}^n \,|\, \left[\begin{smallmatrix} \bar{q}_1 \\ \bar{y}_2 \end{smallmatrix}\right] \in \bar{Q}\right\}$ for which $\bar{\mathcal{I}}_1 x + \bar{q}_1 = \bar{y}_1$. Note that the set $\mathcal{Q}_1$ represents the intersection of the set $\bar{Q}$ with the hyperplane $\left\{\bar{q} = \left[\begin{smallmatrix} \bar{q}_1 \\ \bar{q}_2 \end{smallmatrix}\right] \in \mathbb{R}^m \,|\, \bar{q}_2 = \bar{y}_2\right\}$. Since $\bar{q}$ is uniform on $\bar{\mathcal{Q}}$, it is also uniform on any subset of $\bar{\mathcal{Q}}$, and in particular on this intersection set. Hence, $\bar{q}_1$ is uniformly distributed on $\bar{Q}_1$.

Statement (i) is proved noting that, from (19), an element $x \in \mathcal{I}^{-1}(y) \subset \mathbb{R}^n$ can be written as the mapping of $\bar{q}_1 \in \bar{\mathcal{Q}} \subset \mathbb{R}^n$ through the one-to-one affine transformation $x = \mathcal{I}_1^{-1}(\bar{q}_1 - \bar{y}_1)$. Since bijective linear transformations preserve uniformity [39], it follows that the random variable x is uniformly distributed on $\mathcal{I}^{-1}(y)$.

Point (ii) follows immediately from the fact that the image of a uniform density through a linear operator $\mathbb{R}^n \to \mathbb{R}^s$ with $s \leq n$ is log-concave (see e.g. [39], [50]). □

The result in this theorem can be immediately extended to the more general case when $\mathcal{Q}$ is a compact set. We now introduce an assumption regarding the solution operator $\mathcal{S}$.

Assumption 3 (Regularized solution operator): In the sequel, we assume that the solution operator is regularized, so that $\mathcal{S} = \left[\bar{\mathcal{S}}\ 0_{s,n-s}\right]$, with $\bar{\mathcal{S}} \in \mathbb{R}^{s,s}$.

Remark 4 (On Assumption 3): Note that the assumption is made without loss of generality. Indeed, for any full row rank $\mathcal{S} \in \mathbb{R}^{s,n}$, we introduce the change of variables $T = [T_1\ T_2]$, where T_1 is an orthonormal basis of the column space of $\mathcal{S}^\top$ and T_2 is an orthonormal basis of the null space of $\mathcal{S}$ (in Matlab notation, we write $T_1 = \texttt{orth}(\mathcal{S}^\top)$ and $T_2 = \texttt{null}(\mathcal{S})$). Then, T is orthogonal by definition, and it follows

$$z = \mathcal{S}x = \mathcal{S}TT^\top x = \mathcal{S}\left[T_1\ T_2\right]T^\top x = \left[\mathcal{S}T_1\ \mathcal{S}T_2\right]T^\top x = \left[\bar{\mathcal{S}}\ 0_{s,n-s}\right]\tilde{x} = \widetilde{\mathcal{S}}\tilde{x},$$

where we introduced the new problem element $\tilde{x} \doteq T^\top x$ and the new solution operator $\widetilde{\mathcal{S}} \doteq \mathcal{S}T$. Note that, with this change of variables, equation (2) is rewritten as

$$y = \widetilde{\mathcal{I}}\tilde{x} + q,$$

by introducing the transformed information operator $\widetilde{\mathcal{I}} \doteq \mathcal{I}T$. We observe that any algorithm $\mathcal{A}$, being a mapping from Y to Z, is invariant to this change of variable. It is immediate to conclude that the new problem defined in the variable $\tilde{x}$ and operators $\widetilde{\mathcal{I}}$ and $\widetilde{\mathcal{S}}$ satisfies Assumption 3. ◇

Instrumental to the next developments, we first introduce the degenerate cone (cylinder) in the element space X, with given "center" $z_{\rm c} \in Z$ and radius r, as follows

$$\mathcal{C}(z_{\rm c}, r) \doteq \{x \in \mathbb{R}^n \,|\, \|\mathcal{S}x - z_{\rm c}\| \leq r\} \subset X. \tag{20}$$

Note that this set is the inverse image through $\mathcal{S}$ of the norm-ball $\mathcal{B}(z_{\rm c}, r) \subset Z$. Moreover, due to Assumption 3, the cylinder $\mathcal{C}(z_{\rm c}, r)$ is parallel to the coordinate axes, that is any element x of the cylinder can be written as

$$x \in \mathcal{C}(z_{\rm c}, r) \quad \Leftrightarrow \quad x = \begin{bmatrix} \bar{\mathcal{S}}^{-1}\zeta \\ \xi \end{bmatrix}, \text{ with } \zeta \in \mathcal{B}(z_{\rm c}, r) \subset \mathbb{R}^s \text{ and } \xi \in \mathbb{R}^{n-s}.$$

Hence, for the case $s < n$, the cylinder is unbounded, while for $s = n$ it is simply a linear transformation through $\mathcal{S}^{-1}$ of an ℓ_p norm-ball. Next, for given center $z_{\rm c} \in Z$ and radius $r > 0$, we define the intersection set between the cylinder $\mathcal{C}(z_{\rm c}, r)$ and the consistency set $\mathcal{I}^{-1}(y)$

$$\Phi(z_{\rm c}, r) \doteq \mathcal{I}^{-1}(y) \cap \mathcal{C}(z_{\rm c}, r) \subset X \tag{21}$$

and its volume

$$\phi(z_{\rm c}, r) \doteq \mathrm{vol}\left[\Phi(z_{\rm c}, r)\right]. \tag{22}$$

Finally, we define the set $\mathcal{H}(r)$ of all centers $z_{\rm c} \in \mathbb{R}^s$ for which the intersection set $\Phi(z_{\rm c}, r)$ is non-empty, i.e.

$$\mathcal{H}(r) \doteq \{z_{\rm c} \in \mathbb{R}^s \,|\, \Phi(z_{\rm c}, r) \neq \emptyset\}. \tag{23}$$

Note that, even if the cylinder $\mathcal{C}(z_{\rm c}, r)$ is in general unbounded, the set $\Phi(z_{\rm c}, r)$ is bounded whenever $z_{\rm c} \in \mathcal{H}(r)$, since $\mathcal{I}^{-1}(y)$ is bounded.

We are now ready to state the main theorem of this section, that provides useful properties from the computational point of view of the optimal violation function defined in (17).

Theorem 2: Let $q \sim \mathcal{U}(\mathcal{Q})$ with $\mathcal{Q} \equiv \mathcal{B}(\rho)$, and $\mathcal{S} = \left[\bar{\mathcal{S}}\ 0_{s,n-s}\right]$, with $\bar{\mathcal{S}} \in \mathbb{R}^{s,s}$. Then, the following statements hold

(i) For given $r > 0$, the optimal violation function $v_{\rm o}(r)$ is given by

$$v_{\rm o}(r) = 1 - \frac{\phi_{\rm o}(r)}{\mathrm{vol}\left[\mathcal{I}^{-1}(y)\right]}, \tag{24}$$

where $\phi_{\rm o}(r)$ is the solution of the optimization problem

$$\phi_{\rm o}(r) \doteq \max_{z_{\rm c} \in \mathcal{H}(r)} \phi(z_{\rm c}, r) \tag{25}$$

with $\phi(z_{\rm c}, r)$ and $\mathcal{H}(r)$ defined in (22) and (23), respectively.

(ii) For given $r > 0$, the function (22) is continuous semi-strictly quasi-concave[4] in $z_{\rm c} \in \mathcal{H}(r)$;

[4]A function f defined on a convex set $A \in \mathbb{R}^n$ is semi-strictly quasi-concave if $f(y) < f(\lambda x + (1-\lambda)y)$ holds for any $x, y \in A$ such that $f(x) > f(y)$ and $\lambda \in (0,\, 1)$.

(iii) The function $v_{\rm o}(r)$ is right-continuous and non-increasing for $r > 0$.

Proof: Recalling that $\mu_{\mathcal{Q}}$ is the uniform measure over $\mathcal{Q}$, from (17) we write

$$\begin{aligned} v_{\rm o}(r) &= \inf_{\mathcal{A}} \frac{\mathrm{vol}\left[\left\{x \in \mathcal{I}^{-1}(y) \,|\, \|\mathcal{S}x - \mathcal{A}(y)\| > r\right\}\right]}{\mathrm{vol}\left[\mathcal{I}^{-1}(y)\right]} \\ &= \frac{1}{\mathrm{vol}\left[\mathcal{I}^{-1}(y)\right]} \inf_{z_{\rm c}} \mathrm{vol}\left[\left\{x \in \mathcal{I}^{-1}(y) \,|\, \|\mathcal{S}x - z_{\rm c}\| > r\right\}\right] \\ &= \frac{1}{\mathrm{vol}\left[\mathcal{I}^{-1}(y)\right]} \inf_{z_{\rm c}} \mathrm{vol}\left[\left\{x \in \mathcal{I}^{-1}(y) \,|\, x \notin \mathcal{C}(z_{\rm c}, r)\right\}\right] \\ &= \frac{1}{\mathrm{vol}\left[\mathcal{I}^{-1}(y)\right]} \inf_{z_{\rm c}} \mathrm{vol}\left[\mathcal{I}^{-1}(y) \setminus \mathcal{C}(z_{\rm c}, r)\right]. \end{aligned}$$

Next, we note that this equation can be rewritten as the following maximization problem

$$\begin{aligned} v_{\rm o}(r) &= 1 - \frac{1}{\mathrm{vol}\left[\mathcal{I}^{-1}(y)\right]} \sup_{z_{\rm c}} \mathrm{vol}\left[\mathcal{I}^{-1}(y) \cap \mathcal{C}(z_{\rm c}, r))\right] &(26)\\ &= 1 - \frac{1}{\mathrm{vol}\left[\mathcal{I}^{-1}(y)\right]} \sup_{z_{\rm c}} \phi(z_{\rm c}, r)). &(27) \end{aligned}$$

The statement in (i) follows immediately considering that this optimization problem can be restricted to the set $\mathcal{H}(r)$ where the intersection is non-empty. The existence of a global maximum is guaranteed because $\mathcal{H}(r)$ is convex (and thus compact) and the function $\phi(z_{\rm c}, r))$ is continuous in $z_{\rm c}$.

To prove point (ii), note that problem (25) corresponds to maximizing the volume of the intersection $\Phi(z_{\rm c}, r)$ between the two convex sets $\mathcal{I}^{-1}(y)$ and $\mathcal{C}(z_{\rm c}, r)$. One of them, $\mathcal{I}^{-1}(y)$, is fixed, while the other one is the set obtained translating the cylinder $\mathcal{C}(0, r)$ by $\begin{bmatrix} \bar{\mathcal{S}}^{-1} z_{\rm c} \\ 0 \end{bmatrix}$. Similar problems have been studied in convex analysis, see for instance [57]. In particular, the proof of continuity follows closely the proof of Lemma 4.1 in [20]. That is, consider an arbitrary direction $\xi \in \mathbb{R}^s$, and let $V_{\mathcal{C}}$ be the volume of the set obtained projecting $\mathcal{C}(0, r)$ to the hyperplane normal to ξ. Then, for any $\epsilon > 0$, we have that the difference between the volume of $\Phi(z_{\rm c}, r))$ and $\phi(z_{\rm c} + \epsilon\xi, r))$ is bounded by $\epsilon V_{\mathcal{C}} \|\xi\|$. Hence, $\phi(z_{\rm c}, r)) - \phi(z_{\rm c} + \epsilon\xi, r))$ converges to zero for $\epsilon \to 0$, thus proving continuity.
To prove semi-strict quasi-concavity, consider two points $z_1, z_2 \in \mathcal{H}(r)$ such that $\phi(z_1, r) > \phi(z_2, r)$. Consider then a point $z_\lambda \doteq \lambda z_1 + (1-\lambda) z_2$ where $\lambda \in (0,\, 1)$. From convexity of $\mathcal{H}(r)$ it follows that $z_\lambda \in \mathcal{H}(r)$. Then, the following chain of inequalities hold

$$\begin{aligned} \phi(z_\lambda, r)^{1/n} &= \phi(\lambda z_1 + (1-\lambda) z_2)^{1/n} &(28)\\ &\geq \mathrm{vol}\left[\lambda \Phi(z_1, r) + (1-\lambda)\Phi(z_1, r)\right]^{1/n} &(29)\\ &\geq \lambda \mathrm{vol}\left[\Phi(z_1, r)\right]^{1/n} + (1-\lambda)\mathrm{vol}\left[\Phi(z_1, r)\right]^{1/n} &(30)\\ &= \lambda \phi(z_1, r)^{1/n} + (1-\lambda)\phi(z_1, r)^{1/n} &(31)\\ &> \lambda \phi(z_2, r)^{1/n} + (1-\lambda)\phi(z_1, r)^{1/n} &(32)\\ &= \phi(z_2, r)^{1/n} &(33) \end{aligned}$$

where (29) follows from [57, Theorem 1], (30) follows from the Brunn-Minkowski inequality for convex analysis [42] and (30) follows from the hypothesis that $\phi(z_1, r) > \phi(z_2, r)$. From this chain of inequalities, we have $\phi(z_\lambda, r) > \phi(z_2, r)$, which implies semi-strict quasi-concavity.

To prove point (iii), we note that $v_{\rm o}(r)$ is right continuous and non-increasing if and only if $\phi_{\rm o}(r)$ is upper semi-continuous and non-decreasing. To show upper semi-continuity of the supremum value function $\phi_{\rm o}(r)$, consider the radius $\bar{r} = r_{\rm o}^{\rm wc}(y)$, which is nonzero since $\mathcal{I}^{-1}(y)$ is assumed non-empty. Then, from point (ii) it follows that, for any $\bar{z} \in \mathcal{H}(\bar{r})$, the upper level set $F(\bar{z}) \doteq \{z \in \mathcal{H}(\bar{r}) \,|\, \phi(z, \bar{r}) \geq \phi(\bar{z}, \bar{r})\}$ is strictly convex. Hence, the function $\phi(\cdot, r)$ is quasi-convex, continuous and satisfies the boundedness condition defined in [31]. Then, upper semi-continuity of $\phi_{\rm o}(r)$ follows from direct application of [31, Theorem 2.1]. Finally, to show that $\phi_{\rm o}(r)$ is non-decreasing, take $0 < r_1 < r_2$ and denote z_1 and z_2 be the optimal solutions corresponding to $\phi_{\rm o}(r_1)$ and $\phi_{\rm o}(r_2)$, respectively. It follows that

$$\phi(z_{\rm o1}, r_2) \leq \phi(z_{\rm o2}, r_2) = \phi_{\rm o}(r_2), \tag{34}$$

since $z_{\rm o2}$ is the point where the maximum is attained. On the other hand, from definition (21) and $r_1 < r_2$ we have

$$\Phi(z_{\rm o1}, r_1) = \left(\mathcal{I}^{-1}(y) \cap \mathcal{C}(z_{\rm o1}, r_1)\right) \subseteq \left(\mathcal{I}^{-1}(y) \cap \mathcal{C}(z_{\rm o1}, r_2)\right)$$

and hence

$$\phi_{\rm o}(r_1) = \phi(z_{\rm o1}, r_1) \leq \phi(z_{\rm o1}, r_2). \tag{35}$$

Combining (34) and (35) it follows $\phi_{\rm o}(r_1) \leq \phi_{\rm o}(r_2)$. □

Remark 5 (Unimodality of the function $\phi(z_{\rm c}, r)$): Point (ii) in Theorem 2 is crucial from the computational viewpoint. Indeed, as remarked for instance in [19], a semi-strictly quasi-concave function cannot have local maxima. Roughly speaking, this means that the function $\phi(\cdot, r)$ is unimodal, and therefore any local maximal solution of problem (25) is also a global maximum. Note that from the Brunn-Minkowski inequality it follows that, if there are multiple points $z_{\rm o}(i)$ where $\phi(\cdot)$ achieves its global maximum, then the sets $\Phi(z_{\rm o}^{(i)}, r)$ are all homothetic, see [42]. Further, from the definition of $\Phi(\cdot, r)$, this implies that $\Phi(z_{\rm o}^{(i)}, r) = \Phi(z_{\rm o}^{(j)}, r) + z_{\rm o}^{(i)} - z_{\rm o}^{(j)}$.

This fact is illustrated in Figure 5, where we plot the function $\phi(z_{\rm c}, r)$ for the tutorial problem considered in Example 3, for two different values of r. In the figure on the left, the two sets $\mathcal{I}^{-1}(y)$ and $\mathcal{C}(z_{\rm c}, r)$ always intersect, the function is unimodal, and clearly presents a unique global maximum. In the figure on the right, the radius r is smaller, and there are values of $z_{\rm c}$ for which $\mathcal{C}(z_{\rm c}, r)$ is completely contained in $\mathcal{I}^{-1}(y)$, thus leading to the "flat" region on the top. However, note that this is the only flat region, so that the function is "well-behaved" from an optimization viewpoint. ◇

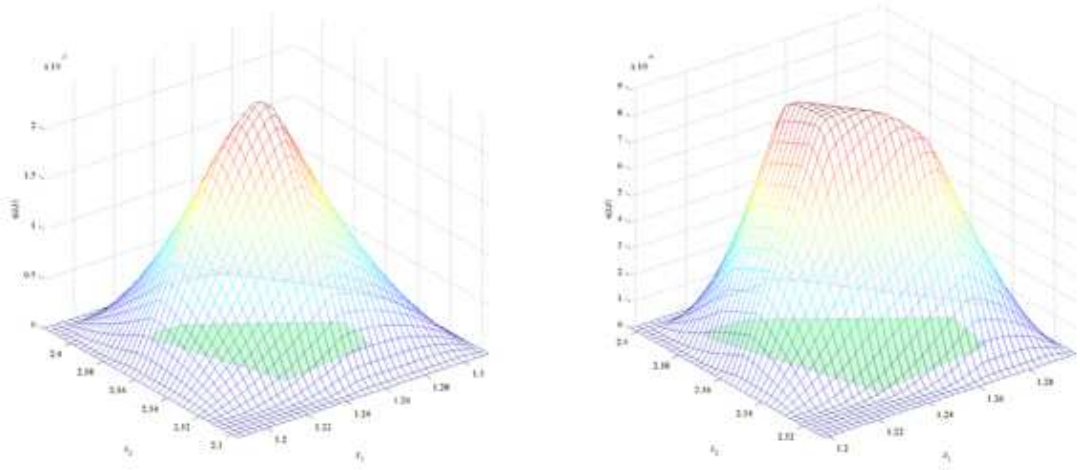

Fig. 5. Function $\phi(z_{\rm c}, r)$ for the tutorial problem considered in Example 3, for $r = 0.0284$ (a) and $r = 0.0150$ (b).

Remark 6 (Probabilistic radius and probabilistic optimal estimator): Theorem 2 provides a way of computing the optimal probabilistic radius of information $r_{\rm o}^{\rm pr}(y, \epsilon)$. Indeed, for given $r > 0$, the probabilistic radius of information (to level ϵ) is given by the solution of the following one-dimensional "inversion" problem

$$r_{\rm o}^{\rm pr}(y, \epsilon) = \min \{r \,|\, v_{\rm o}(r) \leq \epsilon\}\,. \tag{36}$$

Note that point (iii) in Theorem 2 guarantees that such solution always exists for $\epsilon \in (0,\, 1)$, and it is unique. The corresponding optimal estimate is then given by

$$z_{\rm o}^{\rm pr}(\epsilon) = \mathcal{A}_{\rm o}^{\rm pr}(y, \epsilon) = z_{\rm o}(r_{\rm o}^{\rm pr}(y, \epsilon)),$$

where we denoted by $z_{\rm o}(r)$ a solution of the optimization problem (25).

To illustrate, continuing with the tutorial Example 3, we plot in Figure 6(a) the function $v_{\rm o}(r)$ for $r \in (0,\, r_{\rm o}^{\rm wc}]$. We see that $v_{\rm o}(r)$ is indeed non-increasing (actually, it is strictly decreasing), and hence the inverse problem (36) has clearly a unique solution for any $\epsilon \in (0,\, 1)$. $\diamond$

Theorem 2 shows that the problem we are considering is indeed a well-posed one, since it has a unique solution (even though not a unique minimizer in general). However, its solution requires the computation of the volume of the intersection set $\Phi(z_{\rm c}, r)$, which is in general a very hard task. A notable exception in which the probabilistic optimal estimate is immediately computed for q uniformly distributed in $\mathcal{Q}$ is the special case when the consistency set $\mathcal{I}^{-1}(y)$ is centrally symmetric with center $\bar{x}$. Indeed, in this case it can be seen that $\mathcal{S}\mathcal{I}^{-1}(y)$ is also a centrally symmetric around $\bar{z} = \mathcal{S}\bar{x}$, and so is the density $\tilde{\mu}_{\mathcal{S}\mathcal{I}^{-1}}$. Hence, the optimal probabilistic estimate coincides with the center $\bar{z}$,

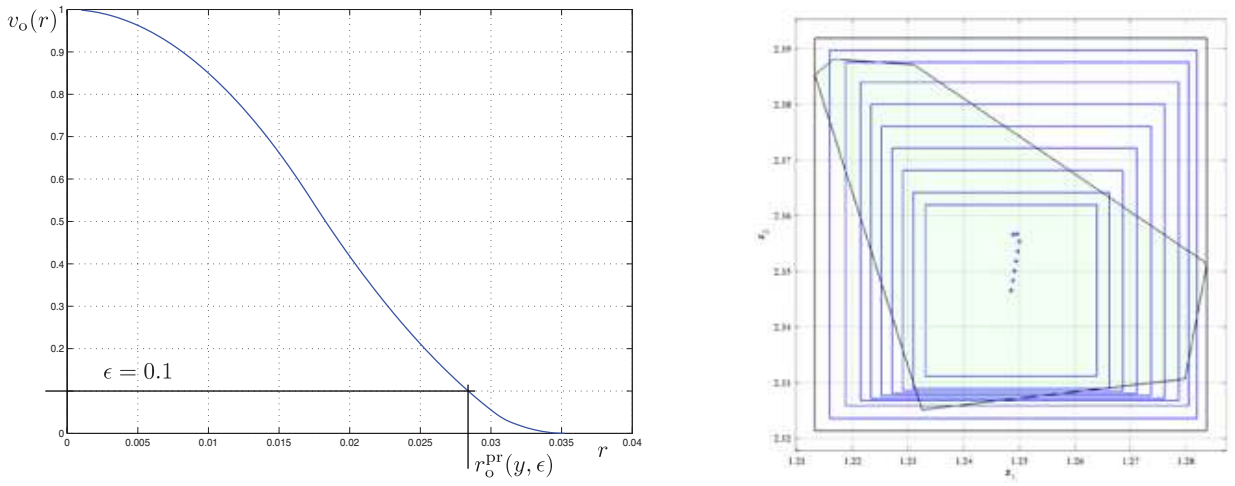

Fig. 6. (a) Plot of the function $v_{\rm o}(r)$ for Example 3, and computation of the optimal probabilistic radius for $\epsilon = 0.1$. (b) Plot of the "optimal" ℓ_∞ balls for different values of r. The crosses denote the corresponding optimal estimates $r_{\rm o}^{\rm pr}(y,\epsilon)$.

since it follows from symmetry that the probability measure of the intersection of $\mathcal{S}\mathcal{I}^{-1}(y)$ with an ℓ_p norm-ball is maximized when the two sets are concentric. Moreover, this estimate coincides with the classical worst-case (central) estimate, which in turn coincides with the classical least squares estimates.

Remark 7 (Weighted ℓ_2 norms): Note that the requirement of $\mathcal{I}^{-1}(y)$ being centrally symmetric is quite demanding in general, but holds naturally when the random uncertainty q is uniformly distributed in a (weighted) ℓ_2 ball of radius ρ, that is

$$\mathcal{Q} = \{q : \|q\|_W \leq \rho\}, \quad \|\xi\|_W \doteq \sqrt{\xi^\top W \xi}, \quad W \succ 0,$$

and $\mu_q(\mathcal{Q}) = \mathcal{U}_q(\mathcal{Q})$. This framework has been also considered in the classical set-membership literature, see for instance [27], and it is well-known that in this case the set $\mathcal{I}^{-1}(y)$ is the ellipsoid

$$\mathcal{I}^{-1}(y) = \left\{ x \in X \,|\, (x - \bar{x}_{\rm ls})^\top \left(\mathcal{I}^\top W \mathcal{I}\right)^{-1} (x - \bar{x}_{\rm ls}) \leq \rho \right\}, \tag{37}$$

centered around the (weighted) least-squares optimal parameter estimate

$$\bar{x}_{\rm ls} \doteq \left(\mathcal{I}^\top W \mathcal{I}\right)^{-1} \mathcal{I}^\top W y.$$

Hence, it follows from symmetry that, for any $\epsilon \in (0,1)$, the probabilistic optimal estimator to level ϵ is given by

$$z_{\rm o}^{\rm pr}(\epsilon) = \mathcal{A}_{\rm ls}(y) = \mathcal{S}\bar{x}_{\rm ls}(y).$$

However, we are not aware of any easy formula to compute the corresponding probabilistic optimal radius $r_{\rm o}^{\rm pr}(y,\epsilon)$, see Section VII for further comments. ◇

VI. Randomized and deterministic algorithms for violation function approximation

In this section, we concentrate on the solution of the optimization problem defined in (25), Theorem 2 for fixed $r > 0$. For simplicity, we restate this problem dropping the subscript from $z_{\rm c}$

$$\textbf{(P-max-int)}: \max_{z \in \mathcal{H}(r)} \phi(z,r), \qquad \phi(z,r) = \mathrm{vol}\left[\mathcal{I}^{-1}(y) \cap \mathcal{C}(z,r)\right]. \tag{38}$$

First, note that this problem is very hard in general. For instance, for ℓ_1 or ℓ_∞ norms, the consistency set $\mathcal{I}^{-1}(y)$ is a polytope and $\mathcal{C}(z,r)$ is a cylinder parallel to the coordinate axes whose cross-section is a polytope. Hence, even evaluating the function $\phi(z,r)$ appearing in (38) amounts at computing the volume of a polytope, and this problem has been shown to be NP-hard in [30].

Remark 8 (Volume oracle and oracle-polynomial-time algorithm): For the case of polytopic sets, the papers [1], [20] study the problem **(P-max-int)** in the hypothetical setting that an oracle exists which satisfies the following property: given $r > 0$ and $z \in \mathcal{H}(r)$, it returns the value of the function $\phi(z,r)$, together with a sub-gradient of it. In this case, in [1] a strongly polynomial-time (*in the number of oracle calls*) algorithm is derived. Note that, even if the problem is NP hard in general, one can compute the volume of a polytope in a reasonable time for considerably complex polytopes in modest (e.g. for $n \leq 10$) dimensions, see [6]. In this particular case, for ℓ_∞ norms, the method proposed by [20] may be used. For instance, for the tutorial Example 3, all relevant quantities have been computed exactly by employing this method. However it should be remarked that, for larger dimensions, the curse of dimensionality makes the problem computationally intractable, and alternative methods need to be devised. ◇

In the next subsections, we develop random and deterministic relaxations of problem **(P-max-int)** which do not suffer from these computational drawbacks.

A. Randomized algorithms for computing *(P-max-int)*

In this section, we propose randomized algorithms based on a probabilistic volume oracle and a stochastic optimization approach for approximately solving problem **(P-max-int)** for generic ℓ_p norms. First of all, we compute a bounded version of the cylinder $\mathcal{C}(z,r)$. To this end, we note that bounds x_i^-, x_i^+ on the variables x_i, $i = s+1, \ldots, n$, can be computed as the solution of the following $2(n-s)$ convex programs,

$$x_i^- = \begin{array}{l} \min\ x_i \\ \text{subject to} \quad x \in \mathcal{I}^{-1}(y) \end{array}, \qquad x_i^+ = \begin{array}{l} \max\ x_i \\ \text{subject to} \quad x \in \mathcal{I}^{-1}(y) \end{array}, \quad i = s+1, \ldots, n. \tag{39}$$

The problems above are convex, and for generic ℓ_p norms can be solved by any gradient-based method. In particular, problem (39) reduces to the solution to $2(n-s)$ linear programs in the case

of ℓ_1 or ℓ_∞ norms. Then, under Assumption 3, we define the cylinder

$$\overline{\mathcal{C}}(z,r) \doteq \left\{ x \in \mathbb{R}^n \mid \left\| \bar{\mathcal{S}} \begin{bmatrix} x_1 \\ \vdots \\ x_s \end{bmatrix} - z \right\| \leq r,\, x_i^- \leq x_i \leq x_i^+, i = s+1, \ldots, n \right\}. \tag{40}$$

Note that the cylinder $\overline{\mathcal{C}}(z,r)$ is bounded, and has volume equal to

$$\mathrm{vol}\left[\overline{\mathcal{C}}(z,r)\right] = \frac{(2r)^s \Gamma^s\left(1/p+1\right)}{|\det(\bar{\mathcal{S}})| \Gamma\left(s/p+1\right)} \prod_{i=s+1}^{n} (x_i^+ - x_i^-) \doteq \mathrm{V}_{\mathcal{C}}. \tag{41}$$

By construction, we have that, for any $r > 0$ and $z \in \mathcal{H}(r)$, $\Phi(z,r) = \mathcal{I}^{-1}(y) \cap \overline{\mathcal{C}}(z,r)$. Note that independent and identically distributed (iid) random samples inside $\overline{\mathcal{C}}(z,r)$ can be easily obtained from iid uniform samples in the ℓ_p-norm ball, whose generation is explained in [8]. Then, a probabilistic approximation of the volume of the intersection $\Phi(z,r)$ may be computed by means of the randomized oracle presented in Algorithm 1, which is based on the uniform generation of iid samples in $\overline{\mathcal{C}}(z,r)$.

Algorithm 1 Probabilistic Volume Oracle

1. RANDOM GENERATION

   Generate N iid uniform samples $\zeta^{(1)}, \ldots, \zeta^{(N)}$ in the s-dimensional ball $\mathcal{B}(z,r)$

   – For $i = 1$ to N

   - Generate s iid scalars according to the unilateral Gamma density $\gamma_j \sim G_{1/p,1}$
   - Construct the vector $\eta \in \mathbb{R}^n$ of components $\eta_j = s_j \gamma_j^{1/p}$, where s_j are iid random signs
   - Let $\zeta^{(i)} = z + r\, w^{1/n} \frac{\eta}{\|\eta\|_p}$ where w is uniform in $[0,1]$

   End for

   Generate N iid uniform samples $\xi^{(1)}, \ldots, \xi^{(N)}$

   – For $i = 1$ to N

   - Generate ξ_j^i uniformly in the interval $[x_{s+j}^-,\, x_{s+j}^+]$, $j = 1, \ldots, n-s$

   End for

   Construct the random samples in $\overline{\mathcal{C}}(z,r)$ as follows

$$\chi^{(i)} = \begin{bmatrix} \bar{\mathcal{S}}^{-1} \zeta^{(i)} \\ \xi^{(i)} \end{bmatrix}, \quad i = 1, \ldots, N$$

2. CONSISTENCY TEST

   – Compute the number of samples inside $\mathcal{I}^{-1}(y)$ as follows

$$N_g = \sum_{i=1}^{N} \mathbb{I}\left(\|\mathcal{I}\, \chi^{(i)} - y\| \leq \rho \right)$$

3. PROBABILISTIC ORACLE Return an approximation of the volume $\phi(z,r)$ as follows

$$\widehat{\phi}_N(z,r) = \frac{N_g}{N} \mathrm{V}_{\mathcal{C}}$$

   where $\mathrm{V}_{\mathcal{C}}$ is defined in (41)

Note that the expected value of the random variable $\widehat{\phi}_N(z,r)$ with respect to the samples $\chi^{(1)},\ldots,\chi^{(N)} \in \overline{\mathcal{C}}(z,r)$ is exactly the volume function $\phi(z,r)$ appearing in **(P-max-int)** that is

$$\mathbb{E}\left[\widehat{\phi}_N(z,r)\right] = \phi(z,r).$$

This immediately follows from the linearity of the expectation

$$\mathbb{E}\left[\widehat{\phi}_N(z,r)\right] = \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N}\mathbb{I}\left(\chi^{(i)} \in \mathcal{I}^{-1}(y)\right)\mathrm{V}_{\mathcal{C}}\right] = \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left[\mathbb{I}\left(\chi^{(i)} \in \mathcal{I}^{-1}(y)\right)\right]\mathrm{V}_{\mathcal{C}}.$$

Then, we have

$$\begin{aligned}\mathbb{E}\left[\mathbb{I}\left(\chi^{(i)} \in \mathcal{I}^{-1}(y)\right)\right] &= 1\cdot\mathrm{Prob}\left\{\chi^{(i)} \in \mathcal{I}^{-1}(y)\right\} + 0\cdot\mathrm{Prob}\left\{\chi^{(i)} \notin \mathcal{I}^{-1}(y)\right\}\\ &= \mathrm{vol}\left[\Phi(z,r)\right]/\mathrm{vol}\left[\bar{\mathcal{C}}(z,r)\right] = \phi(z,r)/\mathrm{V}_{\mathcal{C}}.\end{aligned}$$

Hence, we reformulate the problem at **(P-max-int)** as the following stochastic optimization problem

$$\max_{z\in\mathcal{H}(r)} \mathbb{E}\left[\widehat{\phi}_N(z,r)\right].$$

This problem is classical and different stochastic approximation algorithms have been proposed, see for instance [32], [44] and references therein. In particular, in this paper, we use the SPSA (simultaneous perturbations stochastic approximation) algorithm, first proposed in [46], and further discussed in [48]. Convergence results under different conditions are detailed in the literature, see in particular the paper [24] which applies to non-differentiable functions. This approach is outlined in Algorithm 2.

Algorithm 2 SPSA approach to **(P-max-int)**

1. INITIALIZATION Let $k=0$, choose initial feasible center $z_0 \in \mathcal{H}(r)$ and stepsize sequences $\{a_k\}$, $\{c_k\}$ satisfying $a_k \to 0$, $\sum_k a_k = \infty$, $c_k \to 0$, $\sum_k c_k = \infty$
2. BERNOULLI GENERATION
   - Generate s iid Bernoulli points $\Delta_k \in \{0,\,1\}^s$
   - Define $[\Delta_k^{-1}] \doteq \left[\Delta_{k,1}^{-1}\cdots\Delta_{k,n}^{-1}\right]^{\top}$
3. APPROXIMATE GRADIENT
   - Compute the two perturbed values $\widehat{\phi}_N^{\pm} \doteq \widehat{\phi}_N(z_k \pm c_k\Delta_k, r)$
   - Compute an approximate (sub)gradient as
   $$\widehat{\partial\phi}_N(z_k,r) = [\Delta_k^{-1}]\frac{\widehat{\phi}_N^{+} - \widehat{\phi}_N^{-}}{2c_k}$$
4. SUBGRADIENT ASCENT
   $$z_{k+1} = z_k + a_k\,\widehat{\partial\phi}_N(z_k,r)$$
5. ITERATION Let $k = k+1$ and goto 2

Remark 9 (Scenario-based algorithms): An alternative approach based on randomized methods can be also devised employing results on the scenario optimization method introduced in [7]. In particular, exploiting the results on discarded constraints, see [11], [13], an alternative algorithm can be constructed. The idea is as follows: (i) generate N samples $\chi^{(i)}$ in $\mathcal{I}^{-1}(y)$ according to the induced measure $\tilde{\mu}_{\mathcal{I}^{-1}}$, ii) solve the discarded-constraint random program

$$\min_{z,\gamma} \quad \gamma \tag{42}$$

$$\text{s.t.} \qquad \frac{1}{L}\sum_{i\in I_L} \mathbb{I}\left(\|\mathcal{S}\chi^{(i)} - z)\| \geq \gamma\right) \leq \epsilon \tag{43}$$

where I_L is a set of L indices constructed discarding in a prescribed way $N - L$ indices from the set $1, 2, \ldots, N$. Then, in [11], [13] it is shown how to choose N and the discarded set I_L to guarantee, with a prescribed level of confidence, that the result of optimization problem (42) is a good approximation of the true probabilistic radius $r_{\rm o}^{\rm pr}(y, \epsilon)$. However, this approach entails many technical difficulties, such as the random sample generation in point (i) and the optimal discarding procedure in point (ii), whose detailed analysis goes beyond the scope of this paper, and it is studied in [18]. We also point out that a different approach, also based on scenario optimization and discarded constraints, has been developed in [12] for identification and reliability problems, introducing the concept of interval predictor models. ◇

*B. A semi-definite programming relaxation to **(P-max-int)***

In this section, we propose a deterministic approach to **(P-max-int)** based on a semidefinite relaxation of the problem. We develop our approach focusing on the ℓ_∞ norm; extensions to ℓ_1 and ℓ_2 norms are briefly discussed in Remark 10. First note that, in the case of ℓ_∞ norms, $\mathcal{Q}$ is an hypercube of radius ρ and therefore $\mathcal{I}^{-1}(y)$ is the polytope $\mathcal{P}_X$ defined by the following linear inequalities

$$\mathcal{I}^{-1}(y) = \{x \in \mathbb{R}^n \,|\, \|\mathcal{I}x - y\|_\infty \leq \rho\} \tag{44}$$

$$= \left\{ x \in \mathbb{R}^n \,|\, \begin{bmatrix} \mathcal{I} \\ -\mathcal{I} \end{bmatrix} x \leq \begin{bmatrix} \rho\mathbf{1} + y \\ \rho\mathbf{1} - y \end{bmatrix} \right\} \doteq \mathcal{P}_X \tag{45}$$

where $\mathbf{1}$ is a vector of ones, $\mathbf{1} = [1\ 1 \cdots 1]^\top$. Since the exact computation of the volume of the intersection of two polytopic sets is in general costly and prohibitive in high dimensions, as discussed in Remark 8, we propose to maximize a suitably chosen lower bound of this volume. This lower bound can be computed as the solution of a convex optimization problem. The idea is to construct, for fixed $r > 0$, the maximal volume ellipsoid contained in the intersection $\Phi(z, r)$, which requires

to solve the optimization problem

$$\begin{aligned} \max_{z,x_{\mathcal{E}},P_{\mathcal{E}}} \quad & \mathrm{vol}\left[\mathcal{E}(x_{\mathcal{E}},P_{\mathcal{E}})\right] \\ \text{subject to} \quad & \mathcal{E}(x_{\mathcal{E}},P_{\mathcal{E}}) \subseteq \Phi(z,r), \end{aligned} \tag{46}$$

where the ellipsoid of center $x_{\mathcal{E}}$ and shape matrix $P_{\mathcal{E}}$ is

$$\mathcal{E}(x_{\mathcal{E}},P_{\mathcal{E}}) \doteq \left\{x \in \mathbb{R}^n \,|\, x = x_{\mathcal{E}} + P_{\mathcal{E}} w, \|w\|_2 \leq 1\right\}.$$

The problem of deriving the maximum volume ellipsoid inscribed in a polytope is a well-studied one, and concave reformulations based on linear matrix inequalities (LMI) are possible, see for instance [5], [4]. For completeness, we report this result in the next theorem.

Theorem 3: Let $q \sim \mathcal{U}(\mathcal{Q})$ with $\mathcal{Q} \equiv \mathcal{B}(\rho)$, and $\mathcal{S} = \left[\bar{\mathcal{S}}\ 0_{s,n-s}\right]$, with $\bar{\mathcal{S}} \in \mathbb{R}^{s,s}$. Then, for given $r > 0$, a center that achieves a global optimum for problem (46) can be computed as the solution of the following semi-definite programming (SDP) problem

$$\begin{aligned} z_{\mathrm{o}}^{\mathrm{sdp}}(r) \in \arg_z \min_{z,x_{\mathcal{E}},P_{\mathcal{E}}} & -\log\det P_{\mathcal{E}} \\ \text{subject to } & P_{\mathcal{E}} \succeq 0 \quad \text{and} \\ & \begin{bmatrix} (\rho + e_i^\top(y - \mathcal{I}x_{\mathcal{E}}))I_n & P_{\mathcal{E}}\mathcal{I}^\top e_i \\ \star & \rho + e_i^\top(y - \mathcal{I}x_{\mathcal{E}}) \end{bmatrix} \succeq 0, \\ & i = 1,\ldots,m \qquad (47) \\ & \begin{bmatrix} (\rho - e_i^\top(y - \mathcal{I}x_{\mathcal{E}}))I_n & -P_{\mathcal{E}}\mathcal{I}^\top e_i \\ \star & \rho - e_i^\top(y - \mathcal{I}x_{\mathcal{E}}) \end{bmatrix} \succeq 0, \\ & i = 1,\ldots,m \qquad (48) \\ & \begin{bmatrix} (r + \bar{e}_i^\top(z - \mathcal{S}x_{\mathcal{E}}))I_n & P_{\mathcal{E}}\mathcal{S}^\top \bar{e}_i \\ \star & r + \bar{e}_i^\top(z - \mathcal{S}x_{\mathcal{E}}) \end{bmatrix} \succeq 0, \\ & i = 1,\ldots,s \qquad (49) \\ & \begin{bmatrix} (r - \bar{e}_i^\top(z - \mathcal{S}x_{\mathcal{E}}))I_n & P_{\mathcal{E}}\mathcal{S}^\top \bar{e}_i \\ \star & r - \bar{e}_i^\top(z - \mathcal{S}x_{\mathcal{E}}) \end{bmatrix} \succeq 0, \\ & i = 1,\ldots,s, \qquad (50) \end{aligned}$$

where e_i and $\bar{e}_i$ are elements of the canonical basis of $\mathbb{R}^m$ and $\mathbb{R}^s$, respectively. Moreover, for all $r > 0$, $v_{\mathrm{o}}^{\mathrm{sdp}}(r) \geq v_{\mathrm{o}}(r)$, where we defined

$$v_{\mathrm{o}}^{\mathrm{sdp}}(r) \doteq 1 - \frac{\phi\left(z_{\mathrm{o}}^{\mathrm{sdp}}(r), r\right)}{\mathrm{vol}\left[\mathcal{I}^{-1}(y)\right]}.$$

Proof: The theorem is immediately proved seeing that (47), (48) impose that $\mathcal{E}(x_{\mathcal{E}},P_{\mathcal{E}}) \subseteq \mathcal{I}^{-1}(y)$ while (49), (50) impose that $\mathcal{E}(x_{\mathcal{E}},P_{\mathcal{E}}) \subseteq \mathcal{C}(z,r)$. This problem is an SDP since the equations are linear matrix inequalities in the variables $z, x_{\mathcal{E}}, P_{\mathcal{E}}$, and the cost function is convex in $P_{\mathcal{E}}$. □

From Theorem 3, if follows that the SDP relaxation leads to a suboptimal violation function $v_{\rm o}^{\rm sdp}(r)$.

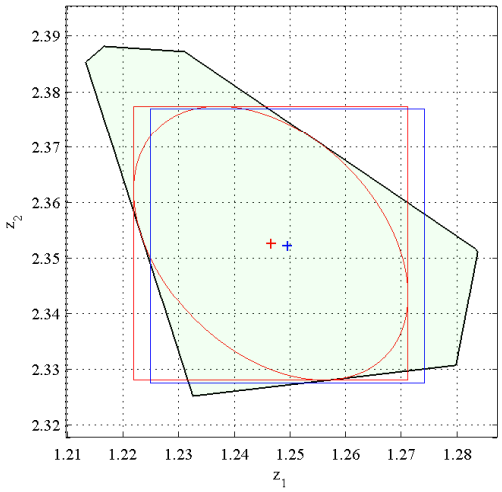

Fig. 7. Optimal ℓ_∞ ball for Example 3, for $r = 0.0284$ (in blue), and optimal ball computed as the solution of the SDP relaxation, with corresponding ellipse $\mathcal{E}(x_\mathcal{E}, P_\mathcal{E})$ (in red).

Remark 10 (SDP relaxations for ℓ_1 and ℓ_2): An approach identical to that proposed in Theorem 3 can be developed for ℓ_1 norm, considering that also in this case the sets $\mathcal{I}^{-1}(y)$ and $\mathcal{C}(z, r)$ are a polytope and a cylinder with polytopic basis. Similarly, an analogous algorithm can be devised for (weighted) ℓ_2 norm. In this case, one should maximize the volume of an ellipsoid contained in the intersection of $\mathcal{I}^{-1}(y)$ and $\mathcal{C}(z, r)$, which are respectively the ellipsoid defined in (37) and a cylinder with spherical basis. It can be easily seen, see e.g. [5], that this latter problem easily rewrites as a convex SDP optimization problem. ◇

VII. Random Uncertainty Normally Distributed and Connections with Least-Squares

In this section, we concentrate on the case when the uncertainty q is normally distributed with mean value $\bar{q}$ and covariance matrix $\Sigma \succ 0$, that is $q \sim \mathcal{N}_{\bar{q},\Sigma}$, and the set $\mathcal{Q}$ coincides with $\mathbb{R}^m$. This permits to draw a bridge between the probabilistic setting introduced in this paper and the classical theory of statistical estimation, which is usually based on normally distributed additive noise. Indeed, it is well known, see e.g. [28], [33] that the minimum variance unbiased estimator for the linear regression model (2) is given by the Gauss-Markov estimator

$$\bar{x}_{\rm ls} = \left(\mathcal{I}^\top \Sigma^{-1} \mathcal{I}\right)^{-1} \mathcal{I}^\top \Sigma^{-1} y,$$

which coincides with the (weighted) least-squares estimator discussed in Remark 7, for $W = \Sigma^{-1}$.

We first remark that this minimum variance problem falls into the average setting of IBC, see [52]. In particular, we recall that the IBC average setting has the objective of minimizing the expected value of the estimation error, that is, for given y, the optimal average radius is defined as

$$r_{\mathrm{o}}^{\mathrm{av}}(y) \doteq \inf_{\mathcal{A}} \left(\mathbb{E}\left[\|\mathcal{S}x - \mathcal{A}(y)\|^2 \right] \right)^{1/2}, \tag{51}$$

and $\mathbb{E}[\cdot]$ denotes the expected value taken with respect to the conditional measure $\tilde{\mu}_{\mathcal{I}^{-1}}$ introduced in Remark 2 (which is also Gaussian, due to well-known properties of normal measures). It follows that the optimal average estimate is immediately given by

$$z_{\mathrm{o}}^{av} = \mathcal{S}\bar{x}_{\mathrm{ls}},$$

for any $y \in Y$. Moreover, in [52, Chapter 6] it is proven that the optimal average radius does not depend on the measurement y, and it can be computed in closed form as

$$r_{\mathrm{o}}^{\mathrm{av}} = r_{\mathrm{o}}^{\mathrm{av}}(y) = \sqrt{\mathrm{Trace}(\mathcal{S}\mathcal{I}^{\top}\Sigma^{-1}\mathcal{I}\mathcal{S}^{\top})}.$$

For what concerns the probabilistic optimal estimate, which is the subject of this paper, we first remark that in the case of normally distributed noise, the definition of the probabilistic radius (13) still applies, observing that the consistency set $\mathcal{I}^{-1}(y)$ defined in (3) in this case is given by

$$\mathcal{I}^{-1}(y) \doteq \{x \in X \mid \text{ there exists } q \in \mathbb{R}^n : y = \mathcal{I}x + q\},$$

and is unbounded. Hence, the "discarded" set $\mathcal{X}_\epsilon$ in (13) can be also unbounded. Note that this is not an issue, since $\tilde{\mu}_{\mathcal{I}^{-1}}$ is defined over all $\mathbb{R}^n$, so that the measure of unbounded sets is well defined.

Similarly to the worst-case and the average settings, the optimality properties of the least-square solution still hold for the probabilistic setting. Indeed, in [52, Chapter 8] it is proven that the optimal probabilistic estimate (to level ϵ) for normal distributions is given by

$$z_{\mathrm{o}}^{pr} = \mathcal{S}\bar{x}_{\mathrm{ls}},$$

for any $y \in Y$. Closed-form solutions for the computation of the probabilistic radius $r_{\mathrm{o}}^{\mathrm{pr}}(\epsilon)$ are not available, and in [52, Chapter 8] the following upper bound is given

$$r_{\mathrm{o}}^{\mathrm{pr}}(\epsilon, y) \leq \sqrt{2 \ln \frac{5}{\epsilon}}\, r_{\mathrm{o}}^{\mathrm{av}}, \qquad \text{for all } y \in Y.$$

However, it is also observed that this bound is essentially sharp when the noise variance is sufficiently small.

VIII. Numerical example

As a numerical example, we consider a randomly generated instance of (2). In particular, $m = 150$ random measurements of an unknown $n = 5$ dimensional vector were generated taking

$$\begin{aligned} \mathcal{I} &= \texttt{round}(20 * \texttt{rand}(\texttt{m}, \texttt{n}) - 10) \\ q &= \rho(2 * \texttt{rand}(\texttt{m}, 1) - 1), \end{aligned}$$

with $\rho = 5$. Similarly, the solution operator was randomly chosen as

$$\mathcal{S} = \begin{bmatrix} -5 & 10 & -7 & 0 & 0 \\ 3 & -4 & 7 & 0 & 0 \\ 2 & 6 & 4 & 0 & 0 \end{bmatrix}.$$

First, the optimal worst-case radius and the corresponding optimal solution have been computed by solving ten linear programs (corresponding to finding the tightest box containing the polytope $\mathcal{S}\mathcal{I}^{-1}(y)$, see [35]). The computed worst-case optimal estimate is $z_{\mathrm{o}}^{\mathrm{wc}} = [-1.831 \ 5.839 \ 11.883]^{\top}$. and the worst-case radius is $r_{\mathrm{o}}^{\mathrm{wc}}(y) = 0.5791$. Subsequently, in order to apply our probabilistic

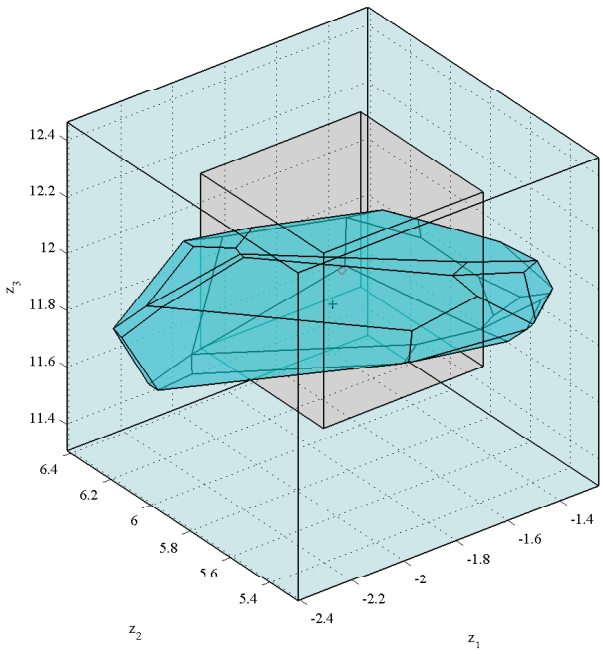

Fig. 8. Consistency set, and optimal "box" discarding a set of measure $\epsilon = 0.1$. The probabilistic optimal radius $r_{\mathrm{o}}^{\mathrm{pr}}(y, 0.1)$ corresponds to the radius of this box. The center, denoted by a star, represents the optimal probabilistic estimate $z_{\mathrm{o}}^{\mathrm{pr}}(y)$.

framework, we fixed the accuracy level to $\epsilon = 0.1$, and computed the probabilistic optimal radius and the corresponding optimal estimate according to definitions (14) and (15). For this size, we were still able to use the techniques discussed in Remark 8 for computing **(P-max-int)** exactly. By

employing a simple bisection search algorithm over $v_{\rm o}(r)$, the probabilistic radius of information was computed as $r_{\rm o}^{\rm pr}(y, 0.1) = 0.3074$. The corresponding optimal probabilistic estimate is given by and $z_{\rm o}^{\rm pr}(0.1) = [-1.773\ 5.869\ 11.969]^\top$. Note that the reduction in terms of radius of information is quite significant, being of the order of 50%. The meaning of our approach is well explained in Figure 8. Indeed, in this figure we see that we look for the optimal "box" discarding a set of probability measure $\epsilon = 0.1$. Note that, in this figure, the volume of the "discarded set" is clearly more than 10% of the total volume. The reason of this is that the probability of the discarded set is measured in the (five dimensional) X space. Figure 9 shows a plot of the violation function $v_o(r)$ computed using the different techniques discussed in this paper. It can be observed that all methods provide very consistent results.

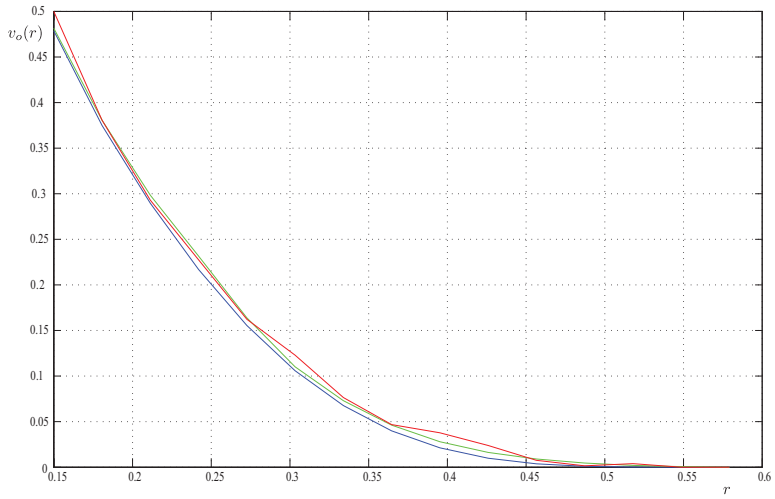

Fig. 9. Function $v_{\rm o}(r)$ for the numerical example. The blue line is the solution obtained computing the volume using the deterministic techniques discussed in Section V, the red line corresponds to the SPSA-algorithm and the green one is the SDP relaxation.

IX. Conclusions

This paper deals with the rapprochement between the stochastic and worst-case settings for system identification. The problem is formulated within the probabilistic setting of information-based complexity, and it is focused on the idea of discarding sets of small measure from the set of deterministic estimates. The paper establishes rigorous optimality properties of a trade-off curve, called *violation function*, which shows how the radius of information decreases as a function of the accuracy. Subsequently, randomized and deterministic algorithms for computing the optimal violation function have been presented. Their performance has been successfully tested on a numerical example.

REFERENCES

[1] H.K. Ahn, S-W. Cheng, and I. Reinbacher. Maximum overlap of convex polytopes under translation. *Algorithms and Computation*, 2010.

[2] E.-W. Bai, H. Cho, R. Tempo, and Y.Y. Ye. Optimization with few violated constraints for linear bounded error parameter estimation. *IEEE Transactions on Automatic Control*, 47:1067–1077, 2002.

[3] E.-W. Bai and K.M. Nagpal. Least squares type algorithms for identification in the presence of modeling uncertainty. *IEEE Transactions on Automatic Control*, 40(4):756–761, 1995.

[4] A. Ben-Tal and A. Nemirovski. Robust convex optimization. *Mathematics of Operations Research*, 23:769–805, 1998.

[5] S. Boyd, L. El Ghaoui, E. Feron, and V. Balakrishnan. *Linear Matrix Inequalities in System and Control Theory*. SIAM, Philadelphia, 1994.

[6] B. Bueler, A. Enge, and K. Fukuda. Exact volume computation for convex polytopes: a practical study. In G. Kalai and G. M. Ziegler, editors, *Polytopes Combinatorics and Computation*, volume 30, pages 131–154. Birkauser, 2000.

[7] G. Calafiore and M.C. Campi. The scenario approach to robust control design. *IEEE Transactions on Automatic Control*, 51(5):742–753, 2006.

[8] G. Calafiore, F. Dabbene, and R. Tempo. Radial and uniform distributions in vector and matrix spaces for probabilistic robustness. In D.E. Miller and L. Qiu, editors, *Topics in Control and its Applications*, pages 17–31. Springer-Verlag, New York, 1999.

[9] G. Calafiore, F. Dabbene, and R. Tempo. Research on probabilistic methods for control system design. *Automatica*, 47:1279–1293, 2011.

[10] G. Calafiore and F. Dabbene (Eds.). *Probabilistic and Randomized Methods for Design under Uncertainty*. Springer-Verlag, London, 2006.

[11] G.C. Calafiore. Random convex programs. *SIAM Journal on Optimization*, 20(6):3427–3464, 2010.

[12] M.C. Campi, G.C. Calafiore, and S. Garatti. Interval predictor models: Identification and reliability. *Automatica*, 45(2):382–392, 1990.

[13] M.C. Campi and S. Garatti. A sampling-and-discarding approach to chance-constrained optimization: feasibility and optimality. *Journal of Optimization Theory and Applications*, 148(2):257–280, 2011 (preliminary version available on Optimization Online, 2008).

[14] M.C. Campi and E. Weyer. Guaranteed non-asymptotic confidence regions in system identification. *Automatica*, 41:1751–1764, 2005.

[15] M.C. Campi and E. Weyer. Non-asymptotic confidence sets for the parameters of linear transfer functions. *IEEE Transactions on Automatic Control*, 55:2708–2720, 2010.

[16] F. Dabbene, M. Sznaier, and R. Tempo. A probabilistic approach to optimal estimation - Part I: Problem formulation adn methodology. *Submitted to the IEEE Conference on Decision and Control*, 2012.

[17] F. Dabbene, M. Sznaier, and R. Tempo. A probabilistic approach to optimal estimation - Part II: Algorithms and applications. *Submitted to the IEEE Conference on Decision and Control*, 2012.

[18] F. Dabbene, M. Sznaier, and R. Tempo. A randomized algorithm for violation function approximation in system identification. *CNR - IEIIT Technical Report*, CNR-IEIIT-12-01-15-TO, 2012.

[19] A. Danilidis and Y. Garcia Ramos. Some remarks on the class of continuous (semi-)strictly quasi convex functions. *Journal of Optimization Theory and Applications*, 133:37–48, 2007.

[20] K. Fukuda and T. Uno. Polynomial-time algorithms for maximizing the intersection volume of polytopes. *Pacific Journal of Optimization*, 2007.

[21] M. Gevers, X. Bombois, B. Codrons, G. Scorletti, and B.D.O. Anderson. Model validation for control and controller validation in a prediction error identification framework - Part I : theory. *Automatica*, 39(3):403–415, 2003.

[22] P.R. Halmos. *Measure Theory*. Springer-Verlag, New York, 1950.

[23] H.D. Hanebeck, J. Horn, and G. Schmidt. On combining statistical and set-theoretic estimation. *Automatica*, 35:1101–1109, 1999.

[24] Y. He, M.C. Fu, and S. Marcus. Convergence of simultaneous perturbation stochastic approximation for nondifferentiable optimization. *IEEE Transactions on Automatic Control*, 48(8):1459–1463, 2003.

[25] A.J. Helmicki, C.A. Jacobson, and C.N. Nett. Control-oriented system identification: A worst-case/deterministic approach in H_∞. *IEEE Transactions on Automatic Control*, 36:1163–1176, 1991.

[26] H. Hjalmarsson. From experiment design to closed loop control. *Automatica*, 41(3):393–438, 2005.

[27] B.Z. Kacewicz, M. Milanese, R. Tempo, and A. Vicino. Optimality of central and projection algorithm. *Systems & Control Letters*, 8:161–171, 1986.

[28] T. Kailath, A. Sayed, and B. Hassibi. *Linear Estimation*. Information and System Science. Prentice Hall, Upper Saddle River, NJ, 2000.

[29] R.L. Karandikar and M. Vidyasagar. Rates of uniform convergence of empirical means with mixing processes. *Stat. and Prob. Let.*, 58:297–307, 2002.

[30] L.G. Khachiyan. Complexity of polytope volume computation. In J. Pach, editor, *New Trends in Discrete and Computational Geometry*, pages 91–101. Springer-Verlag, 1993.

[31] D. Klatte. Lower semicontinuity of the minimum in parametric convex programs. *Journal of Optimization Theory and Applications*, 94:511–517, 1997.

[32] H.J. Kushner and G.G. Yin. *Stochastic Approximation and Recursive Algorithms and Applications*. Springer-Verlag, New York, 2003.

[33] L. Ljung. *System Identification: Theory for the User*. Prentice-Hall, Englewood Cliffs, 1999.

[34] L. Ljung and A. Vicino. Special issue on "system identification" – editorial. *IEEE Transactions on Automatic Control*, 50(10):787–803, 2005.

[35] M. Milanese and R. Tempo. Optimal algorithms theory for robust estimation and prediction. *IEEE Transactions on Automatic Control*, 30:730–738, 1985.

[36] M. Milanese and A. Vicino. Optimal estimation theory for dynamic systems with set membership uncertainty: an overview. *Automatica*, 27(6):997–1009, 1991.

[37] A. Nemirovski and A. Shapiro. Convex approximations of chance constrained programs. *Journal of Optimization Theory and Applications*, 17:969–996, 2006.

[38] B.M. Ninness and G.C. Goodwin. Rapprochement between bounded-error and stochastic estimation theory. *International Journal of Adaptive Control and Signal Processing*, 9:107–132, 1995.

[39] A. Papoulis and S.U. Pillai. *Probability, Random Variables and Stochastic Processes*. McGraw-Hill, New York, 2002.

[40] W. Reinelt, A. Garulli, and L. Ljung. Comparing different approaches to model error modeling in robust identication. *Automatica*, 38(5), 2002.

[41] R.S. Sánchez-Peña and M. Sznaier. *Robust Systems: Theory and Applications*. John Wiley, New York, 1998.

[42] R. Schneider. *Convex bodies: the Brunn-Minkowski theory*. Cambridge University Press, 1993.

[43] J.S. Shamma and K.-Y. Tu. Approximate set–valued observers for nonlinear systems. *IEEE Transactions on Automatic Control*, 5:648–658, 1997.

[44] A. Shapiro. Monte Carlo sampling methods. In A. Rusczyński and A. Shapiro, editors, *Stochastic Programming*, volume 10 of *Handbooks in Operations Research and Management Science*. Elsevier, Amsterdam, 2003.

[45] T. Söderström, P.M.J. Van den Hof, B. Wahlberg, and S. Weiland. Special issue on "data-based modelling and system identification" – editorial. *Automatica*, 41(3), 2005.

[46] J.C. Spall. Multivariate stochastic approximation using a simultaneous perturbation gradient approximation. *IEEE Transactions on Automatic Control*, 37:332–341, 1992.

[47] J.C. Spall. Estimation via Markov Chain Monte Carlo. *IEEE Control Systems Magazine*, 23:34–45, 2003.

[48] J.C. Spall. *Introduction to Stochastic Search and Optimization: Estimation, Simulation, and Control*. Wiley, New York, 2003.

[49] R. Tempo. Robust estimation and filtering in the presence of bounded noise. *IEEE Transactions on Automatic Control*, 33:864–867, 1988.

[50] R. Tempo, G. Calafiore, and F. Dabbene. *Randomized Algorithms for Analysis and Control of Uncertain Systems*. Communications and Control Engineering Series. Springer-Verlag, London, second edition to appear in 2012 edition, 2005.

[51] F. Tjarnstrom and A. Garulli. A mixed probabilistic/bounded-error approach to parameter estimation in the presence of amplitude bounded white noise. In *Proceedings of the IEEE Conference on Decision and Control*, pages 3422–3427, 2002.

[52] J.F. Traub, G.W. Wasilkowski, and H. Woźniakowski. *Information-Based Complexity*. Academic Press, New York, 1988.

[53] J.F. Traub and A.G. Werschulz. *Complexity and Information*. Cambridge University Press, Cambridge, 1998.

[54] M. Vidyasagar and R.L. Karandikar. System identification – a learning theory approach. *Journal of Process Control*, 18:421–430, 2007.

[55] P.G. Voulgaris. On optimal ℓ_∞-ℓ_∞ filtering. *Automatica*, 31(3):489–495, 1995.

[56] E. Weyer. Finite sample properties of system identification of ARX models under mixing conditions. *Automatica*, 36:1291–1299, 2000.

[57] V.A. Zallgaller. Intersection of convex bodies. *Journal of Mathematical Science*, 104:4–7, 2001.